Journal of FST

VOLUME-01 | ISSUE-01 | JULY 2022

ISSN: 2959-4812

Faculty of Science and Technology (FST)

Bangladesh University of Professionals (BUP)

Journal of Faculty of Science and Technology

Volume- 01, Issue- 01, July-2022

ISSN: 2959-4812

Bangladesh University of Professionals (BUP)

Mirpur Cantonment, Dhaka-1216, Bangladesh

Journal of FST

An Academic Journal of Faculty of Science and Technology, Bangladesh University of Professionals (BUP)

Published by

Faculty of Science and Technology
Bangladesh University of Professionals
Mirpur Cantonment, Dhaka-1216
Telephone: +88-02-8000383
Fax: 0088-02-8000443
Email: journal.fst@bup.edu.bd

Correspondence

All correspondence should be addressed to the Chief Editor, Journal of FST, Bangladesh University of Professionals, Mirpur Cantonment, Dhaka-1216

Telephone: 088-02-8000383

Fax: 088-02-8000443

Cover Design

A. K. M. Mohaiminul Islam
Student of BICE 2019

Printed at

Educan International Ltd
House # 31, Road # 4, Flat # 3B
Block # F, Banani, Dhaka-1213
Email: *educan.mail@gmail.com*

Journal of Faculty of Science and Technology

Editorial Board

Message from the Chief Patron

Bismillahir Rahmanir Rahim.

It is a matter of immense pleasure that the inaugural issue of the *Journal of FST*, *volume 1, issue 1*, July 2022 by the Faculty of Science and Technology, Bangladesh University of Professionals (BUP) has come to light. In academia, a journal is perceived to comprise research articles of new knowledge. But there are a wide variety of articles that serve the purpose of communication and provide valuable information to the community. Undoubtedly, university is the best platform where new knowledge and ideas can be created for the betterment of society and thus mankind.

The *Journal of FST is a p*eer reviewed journal determined to maintain high quality comprising contemporary research works focusing on modern day issues, analysis of recent technologies and their performances. I am glad to know that the editorial board has judiciously selected the reviewers on the basis of their scientific proficiency, scholarly figure, rational integrity, and commitment to the journal. An enormous amount of work is done for the development of this journal, and unfeigned effort of all concern is rightly reflected in this inaugural issue. Certainly, we will identify its impact on the science, technology and engineering track.

I would like to congratulate all the authors, reviewers, members of the Editorial Board of the *Journal of FST* at BUP for their highest support in publishing this new journal.

I strongly believe that the *Journal of FST* will contribute a lot in shaping and updating related research and knowledge of the concerned community. Last but not the least, the *Journal of FST* is sincere to contribute in formation of Digital Bangladesh with earnest effort in the field of science and technology.

Major General Md Mahbub-ul Alam, ndc, afwc, psc, MPhil, PhD
Vice Chancellor
Bangladesh University of Professionals

Chief Editor's Note

Faculty of Science and Technology (FST) of Bangladesh University of Professionals (BUP) started its journey in 2012. The faculty is pursuing a paradigm of quality research in the field of modern science and state-of-the art technology. I am immensely pleased to present the maiden issue of the *Journal of FST*, which has been a demand of time.

The *Journal of FST* is a multidisciplinary peer-reviewed annual journal, committed to publish both theoretical and experimental high-quality research papers in the field of Electrical and Electronic Engineering, Computer Science and Engineering, Information and Communication Technology, Environmental Science and Management. It aims to promote the theory and practice of science, technology, and engineering. In addition to research articles, the editorial board welcomes research reports, short communications, book reviews and policy analyses.

For the successful inauguration of this journal, I would like to express my sincere gratitude to our respected Vice-Chancellor, Major General Md Mahbub-ul Alam, ndc, afwc, psc, MPhil, PhD for his continuous support, guidance, inspiration, and cooperation. Moreover, as the Chief Editor, I would like to thank all members of the editorial board along with the associate editors for their continuous dedication and hard work that made possible the first-ever issue of this journal. Initially, we received 47 research articles, which went through rigorous peer review process based on primary screening and anonymous reviewing. Finally, 12 articles have been selected for publication. The acceptance criteria for all papers are originality, quality, aesthetics, and its significance to the journal readership.

Finally, I wish to encourage more contribution from the academic and scientific community to ensure the continued success of the journal. Our commitment is to create a vibrant academic space to explore and publish truly interdisciplinary research resources. We always welcome comments and suggestions from authors, reviewers, and readers for the advancement of this journal through email: *journal.fst@bup.edu.bd.*

Brig Gen Golam Faruque, psc
Chief Editor
&
Dean, Faculty of Science and Technology
Bangladesh University of Professionals

Contents

Contents

Left Blank Intentionally

Analysis of Duplicate Bug Report Detection Techniques

Afrina Khatun[1], Sarker Foysal Mohammud Al Gabid[2], Nazneen Akhter[3], Kazi Abu Taher[4], Tajbia Karim[5]

Abstract

The reporting of large number of duplicate bug reports has generated the need for appropriate duplicate bug report detection techniques. Researchers have developed duplicate bug report detection techniques using different approaches such as Information Retrieval, Machine Learning etc. However, due to rapid development of duplicate detection techniques, it has become difficult to compare and select an appropriate duplicate bug report detection technique. Besides, the usage of different Information Retrieval and Machine Learning techniques have made it more difficult to understand the successes, failures and future opportunities of the proposed techniques. In order to draw a clear picture of the existing techniques developed from the inception to the present, this paper presents a systematic analysis of the duplicate bug report detection techniques. The analysis has been prepared from existing techniques published in ranked conference and journals. The paper has presented insights on the type of input data set used for developing and testing the techniques, the feature selection and pre-processing strategies of bug reports and the type of algorithms and evaluation metrics used for developing the techniques. The paper lastly elaborates the findings established during the discussion of the insights, and presents a road map for future research on the uncovered areas.

Keywords: Duplicate Bug Report Detection, Feature Selection, Machine Learning, Information Retrieval, Neural Network, Deep Learning.

---

[1] Lecturer, Department of Information and Communication Technology, Bangladesh University of Professionals (BUP), Email: afrina.khatun@bup.edu.bd

[2] BICE Student, Department of Information and Communication Technology, Bangladesh University of Professionals (BUP), Email: soykot070917@gmail.com

[3] Lecturer, Department of Computer Science and Engineering, Bangladesh University of Professionals (BUP), Email: nazneen.akhter@bup.edu.bd

[4] Professor, Department of Information and Communication Technology, Bangladesh University of Professionals (BUP), Email: kataher@bup.edu.bd

[5] Assistant Professor, Department of Information and Communication Technology, Bangladesh University of Professionals (BUP), Email: tajbia.karim@bup.edu.bd

1. Introduction

Due to the changing nature of software systems, a bug in the source code of software is common to all software systems. These bugs are submitted in report formats for resolution. A reported bug report logically represents the information of a source code bug. However, in open source systems, large number of developers work from different locations in a single project. Therefore, it is very common that a similar bug report can be submitted by a number of developers. Besides, in open source systems, the users also have the privilege to report bugs. As a result, same bug can be reported repetitively which creates duplicate bug reports. For example 30% of Firefox's reported bug reports are duplicate of already reported bugs (Ebrahimi et al., 2019). Duplicate bug reports add cost to software development and maintenance process in terms of time and effort (Rahman et al., 2020). The identification and resolution of duplicate bugs consume the efforts of bug triggers and software developers which can be spent on unique bug reports. To reduce the cost, a number of techniques for identifying duplicate bug reports have been proposed in literature. Anvik et al. (2005) first characterized the problem of duplicate bug reports in open-source systems. The paper analyzed the bug reports from two open-source projects such as Eclipse and Mozilla Firefox. The analysis characterizes the bug reports from four different perspectives such as - the type of bug reports, the rate of bug reporting, the duration of bug resolution and the role of bug resolver.

In order to conduct the systematic analysis, the literature is searched in a step by step approach. At first, a google scholar search has been conducted using the query string "Duplicate Bug Report Detection". However, the google search technique returns all matching results that have the query string terms in any segment of the searched articles. If the google scholar search is performed by setting the option named "sort by relevance" as checked, then the false positive results occur at the end of the returned search list. Therefore, the search is performed by setting the option "sort by relevance" checked for each of the years from 2017 to 2021. For each year, the top most 100 articles have been identified. Therefore, a total of 500 articles have been collected for processing in the next phase. During the second phase, the articles which are non-English, database files, books or letter type articles are excluded. Later, the title and abstract of the full text articles are checked for relevancy with the searched query string. Out of the 500 identified articles from year 2017 to 2021, total 74 articles are found to propose different techniques of duplicate bug report identification. In the last phase, the ranking of the conferences and journals where the 74 articles are published is checked using Computing Research and Education (CORE) site (Education, 2021). Out of the 74 articles, 23

articles are published in ranked conferences and journals. Therefore, these total 23 articles are selected for the final systematic analysis.

For analyzing the existing techniques and identifying the future scope in duplicate bug detection, this paper tends to answer three Research Questions (RQs).

- RQ1 demonstrates the input data types used for duplicate bug report detection.
- RQ2 identifies the bug report properties which are used by existing duplicate bug detection approaches.
- RQ3 discusses about the algorithms which are being used for duplicate bug report detection.

The rest of the paper is organized into following sections. Section 2 shows the general model of a duplicate bug report detection technique. Section 3 discusses the existing duplicate detection techniques. Section 4 answers RQ1 by enlisting the input data sets used for developing and assessing the existing techniques. The selection and pre-processing strategies of bug report features is discussed in Section 5. This section also answers RQ2 by identifying the most used bug report features. Section 6 answers RQ3 by describing the machine learning, information retrieval and deep learning approaches used for developing the existing approaches. Finally, section 7 concludes the paper by summarizing the findings and pointing the future research scopes in the field of duplicate bug detection.

2. General Model of Duplicate Bug Report Detection Techniques

The overall duplicate bug detection techniques consist of some common sequential steps. Figure 1 represents the general model of duplicate bug report detection technique which consists of *Feature Selection*, *Bug Report Pre-processing* and *Learning Algorithm or Model*. The bug reporting systems such as Bugzilla, Jira etc. enable the developers and users to report bugs using customer friendly software interface. The tracking systems collect and maintain a variety of information about the reported bugs as bug properties (Serrano & Ciordia, 2005). These properties include - bug report id, assignee information, reporting date, component details, product details, bug severity, bug tossing and management history, bug status, bug severity, developer's comments during bug resolution, bug summary, bug description and many more. Some secondary properties such as bug screen shots, bug stack trace, link to source code commit messages etc. are also included sometimes. Each of the property represent different information regarding the bug. The Feature Selection phase selects appropriate properties from the bug report for learning the algorithm or model in the latter phases.

The properties of bug report are written in different formats such as natural language, categorical format etc. Natural language text may contain redundant information which do not represent the actual characteristics of the reported bug. Therefore, the Pre-processing phase processes the content value of bug report properties. Tokenization, stemming, stop word removal are some of the known text pre-processing techniques used by existing techniques. Each property of the bug report represent different information of the reported bug. Finally, in the Learning Algorithm or Model a model is prepared using either machine learning, deep learning or information retrieval based techniques. The features are fed into these models identifying the duplicates. For machine or deep learning based approaches, a bug report is generally classified as duplicate or not. On the other hand, information retrieval based techniques return a ranked list of bug reports suggested to be duplicates.

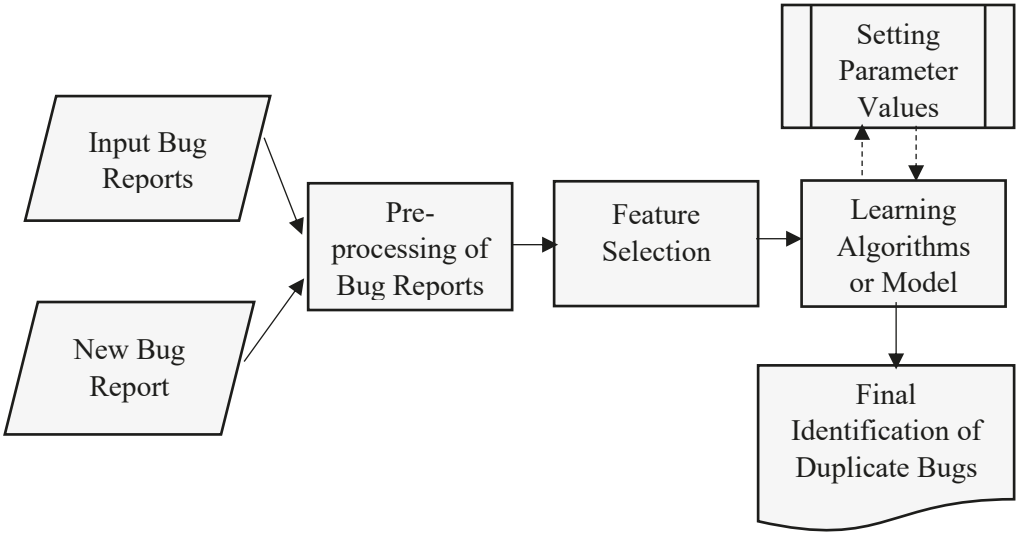


Figure 1: Components of General Duplicate Bug Report Detection Techniques

3. Literature Review

With the increasing number of duplicate reports in the bug repository, the need for duplicate bug detection techniques becomes inevitable. In this essence, a number of duplicate detection approaches have been devised. The existing techniques differ from each other in using bug report properties, applying preprocessing techniques, implementing algorithms and evaluating the results. Therefore, this section describes the existing duplicate bug report detection techniques.

Jalbert and Weimer (2008) devised a duplicate bug report classification technique that classifies the bug report as duplicate as the report arrives in the system. The technique comprises of three steps which are surface (severity, OS version and number of source patch or screenshot attachment) and textual (title and description) feature extraction, text similarity measurement and graph clustering. The similarity value of textual features are calculated using vector dot product. A graph is then induced where the nodes represent bug reports and edges represent their calculated similarity value. Next, the existing reports are clustered using graph clustering algorithm. On arrival of a new report, the textual similarity is checked and the report is classified in the clusters for duplicate identification. The experiment includes a representation of a real-time bug reporting environment using 29,000 bug reports from Mozilla project. The proposed approach achieved around 52% recall rate in identifying duplicates, however the rate seems poor in comparison with other existing techniques.

Johannes and Mira (2013) proposed an approach that applies stack traces to machine learning algorithms for detecting bug report duplicates. However, the technique may fail if source code attachments are missing in bug reports. Neysiani and Babamir (2019a) also presented a duplicate detection approach which combines both IR and ML based algorithms. The paper shows that when combined the IR based technique shows similar performance as ML based techniques. Hence, ML based models provide better performance in duplicate detection than the IR based models alone.

Tian et al. (2016) developed a unified model for bug report assignee recommendation by combining developer's activity information and bug localizing information. The proposed model extracts the title and summary text from the bug reports, and the comments and identifiers from the source code files. These extracted text then go through various text processing steps such as - tokenization, stemming and stop word removal. The model also creates a list of developer profiles which represent the list of bug reports and source code files worked on by the corresponding developers. Next, to train the assignee recommendation model sixteen (16) different features are proposed and extracted using the processed texts of bug reports and source code files. Among the sixteen features, the first five (5) features indicate the similarity between the new bug report and the source code files added, edited or deleted by a developer. In order to calculate the similarity value, cosine similarity and BM25 technique is applied. The features six (6) to ten (10) extract the similarity between the new and the

previously fixed bug reports by a developer. Feature eleven (11) and twelve (12) represents the frequency and recency of a developer's bug resolution. Features 13 to 16 enables the model to incorporate the location information of bug reports and developer commits. The similarity between the new bug report's buggy code and developer's previously fixed source code is measured using the cosine similarity and BM25 technique, for forming features thirteen (13) and fourteen (14). Lastly, feature fifteen (15) and sixteen (16) checks if a developer has ever touched the source code files corresponding to the new bug report. When a new bug report arrives, the model recommends a ranked developer's list by calculating a weighted sum of k features selected from the 16 features (Lee & Lin, 2014). More features can be added, as the work of Neysiani et al. (2020) shows that consideration of product feature plays vital role. Besides, the k selected features for the three data sets show that the features from activity information play better role than location based ones.

Ebrahimi et al. (2019) devised a duplicate bug report detection technique which leverages the stack traces of bug reports. The bug report properties such as product, component and status is examined to identify the master bug report and its corresponding duplicates. Then, the identified bug reports are used to create Duplicate bug report Groups (DG) where each DG contains the stack trace of the master bug and its corresponding duplicates. In this case the bug reports having minimum four stack traces are taken into consideration. Next, the bug report of each DG is used to train a Hidden Markov Model (HMM). The stack traces of bug reports are treated as hidden state sequences and the DG groups are treated as observations in the HMM training. Baum-Welch algorithm is used to estimate the three parameters of HMM. For each DG group 60%, 10% and 20% bug reports are used for training, validation and testing of the HMM model respectively. When a new bug report arrives, the stack trace of the report is collected and matched with each of the trained HMM to identify the probability of matching score. A highest probability score represents the highest probability of being duplicate of the existing DG groups. The technique will fail for bug reports which does not contain any stack traces.

Neysiani and Babamir (2019b) presented a Longest Common Sequence (LCS) based new feature for measuring similarity between two bug reports. From each bug report identical, categorical, textual, and contextual features are elicited and converted into a sequence of features. The textual and contextual features are measured using BM25F and cosine similarity respectively. Next, the longest common sequence is calculated between the bug reports for identifying duplicates.

The proposed technique achieves a high accuracy, precision and recall rate of 96.09%, 98.43% and 97.27% respectively. The paper added a new feature with the existing ones, however, finding the appropriate feature for calculating similarity between bug reports is still an unanswered research question.

Neysiani and Babamir (2019c) proposed the individual Manhattan distance comparison method as an alternative to cosine space similarity for each subject in contextual features (Neysiani & Babamir, 2019c). The study is an extension to previous papers with using the different similarity metric. The proposed model was trained by decision tree and linear regression classifiers and achieved accuracy 96.65% and 81% respectively and recall rate 95.89% and 73.18% respectively. Wang et al. devised an duplicate detection approach which considers the natural language info and source code execution info. The approach first calculates the Natural Language based Similarities (NL-S) between the new and existing bug reports. Next, the Execution information based Similarities (E-S) are measured between the new and remaining bug reports. Finally, reports are selected using comparisons based on two heuristics, where the first is to merge the E-S and the NL-S into one joint comparison and the second is to try to identify whether the natural language or the implementation info is the main issue in perceiving the duplicate reports. The technique used TF-IDF similarity measure for checking equality which tends to achieve poor accuracy in comparison with machine learning approaches.

Poddar et al. (2019) proposed a neural architecture to identify duplicate reports by considering the latent issues of the reports. The paper applied IR based topic modeling for identifying the bug reports. However, the paper mentioned that bug report written by general users and experienced developer differ from each other in terms of technical word selection. Therefore, considering the bug reports separately based on their submitter can also affect the performance of duplicate detection. Soleimani and Morteza (2020) presented an approach to estimate the impact of typo on Duplicate Bug Report Detection (DBRD). The evaluation of the approach on the Android dataset shows that the typos improvement can increase the accuracy and recall of DBRD at most 1% in average, which is trivial. The proposed technique also used the textual, categorical, contextual and temporal features. The results indicated that removal of typos cannot improve the performance of existing techniques.

Jianjun et. al (2020) proposed a duplicate bug detection technique using Dual-Channel Convolutional neural network. First, the technique extracts structured

(such as component, product, priority) and unstructured (such as summary, description) information from the bug report and converts them into a text document. Next, the text documents go through pre-processing steps which include tokenization, stemming, stop word removal and case conversion. After pre-processing, the words of the bug reports are converted into a corpus using word2vec model. As a result, each word is represented using a single dimension vector and each bug report is represented using a two dimension matrix containing the vectors. The duplicate bug report pairs are then represented by combining the single-channel bug report matrices into dual-channel bug report pairs. Next, the training phase fed the dual-channel bug report matrices into the Convolutional Neural Network (CNN) for training the model. The similarity score is compared with a threshold value to classify the bug report pairs as duplicates. The proposed technique considered only 300 words from each bug reports which may ignore important information of the remaining part of the bug reports. Besides the technique ignored the source code attachment or files while consideration of the bug report properties.

Neysiani et al. (2020) proposed an efficient feature extraction model for duplicate bug report detection. The technique starts with pre-processing the existing bug reports by null value detection, homologize bug report formats, tokenization of keywords etc. After pre-processing, the duplicate and non-duplicate bug reports are listed and inputted into the feature extraction phase. For extracting efficient features from bug reports, this phase considers four types of features which are textual, temporal, categorical and contextual. The textual feature are extracted using TFIDF, BM25F, Longest Common Sub-sequence (LCS) and aggregated functions such as maximum, minimum and average values of conventional TF and IDF values in uni-gram and bi-gram forms. The temporal features are extracted using the interval between bug report ids and opening dates. The categorical features are collected by comparing the similarity of product, component, priority, type and version information of bug reports. Contextual features are elicited by calculating the cosine similarity between contents of bug reports. Next, the efficiency of each extracted feature is checked using a Efficiency Detector Value (EDV). The EDV value is calculated using a new heuristic approach that takes the weighted average of the normalized information gain ratio, Gini index, chi-square, Principal Component Analysis and correlation of each feature. Using the extracted features a duplicate bug report detector model is generated. Although the model takes only 5 minutes to predict if a bug report is duplicate, the model compares each new bug with every existing bug report which requires huge computation.

The paper also mentioned that the contextual features alone cannot predict the duplicate bug report effectively.

Alkhazi et al. (2020) proposed an extended version of Tian et. al by adding four more features which can be used in a ranking model. The proposed technique adds four features for calculating the similarity between the new bug report and previously committed messages by developers. The features are collected from bug report commit messages and source code API documentation. Finally, total twenty two (22) features were extracted to identify the appropriate feature combination for training the bug report assignment model. Naive aggregation, Ordinal regression and Learning-to-rank are used to combine the extracted features. Sometimes only the experienced developers have permission to commit, but the resolution may be done by other developers, therefore detail of all commit messages need to be considered.

Jiang et al. (2020) developed a security bug report detection model known as LTRWES, by combining learning to rank with word embedding. LTRWES detects security related bug reports using four steps which are - Learning to Rank, Selecting, Training and Predicting. The Learning to Rank phase starts with ranking the NSBRs based on their similarity with the SBRs for correctly labeling all the bug reports. In this regard, the summary and description fields of each bug report is extracted and pre-processed. The pre-processing step includes text tokenization, lowercasing letters, stop word removal and stemming. Next, the similarity score between a pair of NSBR and SBR is calculated using BM25Fext technique. The similarity score is represented using a matrix where the rows indicate NSBRs and columns indicate SBRs. The average of the similarity scores in a row represents the actual similarity score of a NSBR with respect to the other SBRs. The NSBRs are then ranked based on their average similarity scores where the ranking shows the lower similarity scored bug reports at the top. The selection phase identifies the appropriate NSBRs by applying either of Multiple Selector (ms-selector) or Roulette Wheel Selection (rs-selector) algorithm on the previously ranked NSBRs. The ms-selector algorithm selects the top K lower scored bug reports from the ranked NSBRs. On the other hand, re-selector algorithm also selects K bug reports based on the probability of dissimilarity between the NSBRs and SBRs. The value of K is specified by multiplying the ratio value with the number of SBRs. Next, the top K selected NSBRs and all the SBRs are feed to the prediction model for training. The training phase first creates a vector representation of each bug report using the Continuous Bag of Words (CBOW) model. Naive Bayes, Multilayer Perceptron, Random Forest, K- Nearest Neighbor (KNN), Logistic Regression,

and Support Vector Machine (SVM) techniques are used to train a model using the bug report vectors. Finally, in the prediction phase, when a new bug report comes, the bug report is also pre-processed and converted into a vector representation using the CBOW model. The model takes the vector representation as input and predicts the bug report as NSBR or SBR. The model is tested both on within project and cross project. But while testing cross projects, only one project was taken as cross project. For multiple cross projects how the technique performs can be analyzed further.

Akilan et al. (2020) proposed a computational efficient double tier duplicate bug detection system. The overall technique is divided into two phases – clustering and classification. For each bug report, structured (component, product, priority) and unstructured (summary, description) information's are extracted. The clustering phase prepossesses the bug reports and removes the redundant information. Next, the bug reports are clustered using Latent Dirchilect Allocation (LDA) topic modeling. When a new bug report comes, the Top N clusters which have similarity with the bug report are extracted. The selection of Top N clusters reduces the necessity of matching a bug report to all existing reports. As a result computational efficiency is achieved. In the classification phase, the bug reports are represented in vector formats using Word2Vec, GloVe and FastText. The similarity between these vectors are calculated using cosine similarity and Euclidean similarity. The proposed technique achieved 67% recall rate with 3 times less computation. However, the recall rate of topic modeling based techniques seems lower than exiting machine and deep learning techniques.

Kumar et al. (2020) developed a classification technique for identifying duplicate bugs. The technique first extracts the categorical (product, component and version) and textual (headline and description) features and preprocesses those. In feature generation phase, three types of features have been calculated such as text statistical, semantic and contextual feature. These features are fed to train the machine learning classification which uses the XGBoost algorithm. When new bug report arrives three servers known as App server, Model server and Embedding server works together to get the new report as input and classify the report as duplicate. The developed model was tested on Mozilla and Cisco project which achieved 90%, 98%, 94% and 87% precision, recall, F1-score and accuracy respectively. Although the paper used textual, categorical and customized extracted features, it did not mention which feature have highest impact in identifying appropriate duplicates.

Kukkar et al. (2020) proposed an automatic approach for detecting and classifying duplicate reports based on deep learning. The proposed approach considered the textual features such as title, summery and description and identical features such as bug id. Deep learning-based model mainly CNN is applied for eliciting the word's semantic and morphological relationship for the textual similarity assessment between bug reports. The proposed model achieved higher accuracy rate in between 85% - 99% and recall rate is 79%-94%. The proposed technique did not consider the contextual and temporal features which could impact the performance of CNN feature extraction.

By combining the attention based and context based feature, a duplicate bug report detection has been proposed by Rocha and Carvalho (2021). The overall technique is divided into three phases – training, retrieval and classification. In training phase, the bug report textual keywords are weighted for topic extraction and one hot encoding is created for contextual features. Next, a quotient loss function is devised for calculating the similarity between bug report embedding. The technique has been applied on Eclipse, NetBeans and Mozilla database which achieved accuracy of 84% accuracy. The application of the proposed loss function on closed or industrial project can be another future research scope.

Mahfoodh and Hammad (2022) proposed a duplicate bug report detection technique for predicting the risk factor of software components. The techniques uses the title / summary and description of bug report and extracts the word tokens. Next, the word tokens are converted into array which are fed into neural network. The similarity is measured using two approaches. The first similarity approach iterate on the words of one bug report to find its similar word with another bug report using Euclidean distance. On the other hand, the second similarity measure iterate on the words however within a given range. The technique achieved an average precision of 99.89%. The proposed techniques considers a fixed given range value for checking words. The increasing number of range value may affect the computational performance of the technique.

Few existing papers have also presented survey of the above mentioned duplicate bug report detection techniques (Neysiani et al., 2019a). Most of the papers only analyzed the different algorithms and evaluation metrics used for developing the duplicate detection techniques. However, none of the existing papers analyzed the effects of datasets, pre-processing techniques and bug report features while developing the detection techniques.

The difference in considering different datasets, pre-processing techniques, bug report features, algorithms and evaluation metrics raises the issue of generality of existing duplicate report detection approaches. Therefore, the detailed analysis regarding selection of input datasets, pre-processing mechanisms, features, algorithms and evaluation metrics may uncover the areas of improvement for future research.

4. Input Data Set

Open source systems are developed with contribution of developers from various location of the earth. Besides, now-a-days most of the software systems collect feedback from users to provide continuous support and maintenance. As a result, software bugs are reported by developers, testers and also users all over the world. To track and maintain these huge load of information, the open source software systems use bug tracking repositories such as – Jira, Bugzilla etc. Existing papers have used bug data set from the open source bug tracking systems (Ebrahimi et al., 2019; He et al., 2020; Neysiani et al., 2020; Tian et al., 2016). Some techniques have also incorporated software source codes along with the bug reports for better identification of duplicate bugs.

RQ1: What type of data are being used for duplicate bug report detection?

The most common data used for bug assignment, localization, classification and duplicate identification is bug repository. A number of open source bug repositories are available now-a-days. The bug repositories provide a number of functionalities to easily search and view the bug reports. Besides, the bug tracking systems have also provided different end-user and server-side utilities as third party extensions for easy and public access to submitted bug reports. Therefore, almost all of the existing techniques have used the open source bug tracking systems to collect input data. Many open source software projects uses these bug tracking systems for maintaining their bug repositories.

Figure 2 have listed the open source projects that have been used by existing duplicate bug detection techniques. The figure depicts that open source systems which maintain the bug repository consistently are used by most of the techniques. For example Eclipse is used by most of the existing techniques as a data set. Eclipse bug repository is maintained using Bugzilla. Bug reports can be searched and extracted in different formats. For example - JDT, SWT, ANT, UI are some of the products of Eclipse. Tian et al. (2016) and Alkhazi et al. (2022) have applied their proposed techniques on Eclipse JDT, SWT and UI product specific bug reports for evaluation.

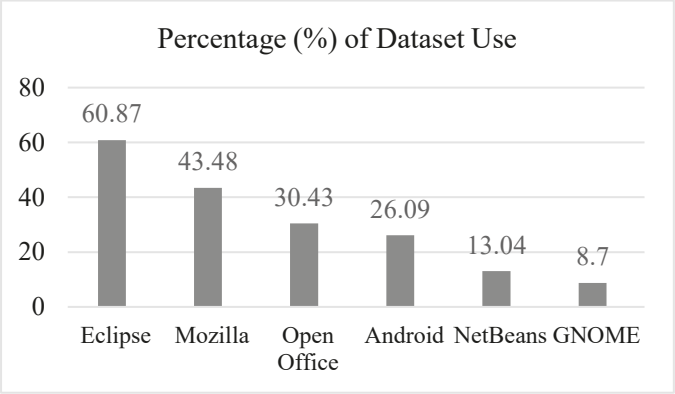


Figure 2: The usage of bug dataset by existing techniques

Besides Eclipse dataset, the Mozilla, Open Office, Android, NetBeans and GNOME are one of the most used dataset by most of the existing techniques (Ebrahimi et al., 2019), (He et al., 2020), (Sabor et al., 2017). Bugzilla has provided different plugins and websites for managing bug data of projects such as Bugzilla GNOME, Bugzilla Mozilla etc. (Bugzilla, 2021). Yuan et al. has collected security specific bug reports from Chromium, Derby, Camel, Wicket and Ambari dataset. In order to download the reports an open source web crawler known as Scarpy has been used (Zou et al., 2018). The tool is implemented in python for specifically crawling security specific bug reports using multi-type feature analysis.

For learning a model substantial amount of fixed bug reports are required. Yuan et al. collected minimum of 1000 bug reports from Ambari project for developing a security bug detection model (Jiang et al., 2020). The reason of selecting a lower number of bug reports is the availability of security bug reports. On the other hands, most of the techniques have used more than 10,000 bug reports as data set. Xiao et al. collected 2,73,710 bug reports from Eclipse bug repository for developing a heterogeneous information network to detect duplicate bugs (Xiao et al., 2020). The bug tracking systems enable the easy submission and maintenance of large scale bug reports. Therefore, it is evident that most of the techniques have been developed and tested on data set which are available in open source bug tracking systems and which already have large scale bug data set.

For ensuring the effectiveness, a duplicate bug report detection technique needs to be tested on both open source and closed source projects. The process of bug detection, bug reporting, developer involvement for bug resolution etc. differs between open source and closed source projects. Only a few techniques have

applied closed source projects for development and evaluation (Cooper et al., 2021a, 2021b).

Bug reports in open source systems are reported using bug tracking systems. Each bug tracking system provides a predefined form structure for reporting bug reports. Bugs are reported using these forms and can be downloaded using plugins provided by the bug tracking systems (Bugzilla, 2021). By using the plugins the bug reports can be downloaded in XML, CSV and plain text format. Figure 3 shows the sample Eclipse bug report having id 519169 in XML format.

```
<bug>
        <id>519169</id>
        <creation_ts>2017-07-04 10:54:44 -0400</creation_ts>
        <summary>Vulnerability found in Eclipse</summary>
        <product>andmore</product>
        <component>Core</component>
        <version>0.5.0</version>
        <status>ASSIGNED</status>
        <priority>P1</priority>
        <severity>critical</severity>
        <long_desc><commentid>2849292</commentid>
                <who name="Alon Boxiner">alonbo</who>
            <thetext>Steps to reproduce: 1. Open a new Android
                                                project.</thetext>
        </long_desc>
</bug>
```

Figure 3: Sample eclipse bug report in XML format

Few approaches have used video-based bug reports as input (Cooper et al., 2021a, 2021b). These approaches leveraged screen-recording features of Android and iOS device for capturing the video of bugs. The videos showing error are reported as bugs for further resolution. The videos are converted into a consecutive series of images. Next, the text of the images are extracted and matched with incoming bug reports for duplicate detection.

The source code repository are collected from version control systems such as GitHub, Bit bucket etc. Each version control system provides command line functionalities to download the repository. Besides, the associated commit messages can also be extracted using the terminal commands. The comments of commit messages can be extracted in plain text, XML etc. formats. The change history of source code files can also be extracted from commit details.

The above discussion answers the RQ1 by identifying that most of the existing techniques generally use fixed bug reports as input data set. Besides, open source systems which have maintained rich bug report repository (such as Eclipse, Mozilla, NetBeans etc.) for long time are also preferred by most of the techniques. Along with the bug reports, the source code, commit files, version history and comments are also used as input data. Now-a-days the video and screenshot based bug reports are also being used. Thus the study of video or screenshot based duplicate bug detection can be explored by researchers. Although most of the technique has used open source bug repository, only a few approaches have tested the technique on closed source projects. Therefore, future research scope lies in finding appropriate duplicate detection technique for closed source and cross projects.

5. Feature Selection and Pre-processing

After collecting the input bug reports, appropriate bug report property or feature needs to be selected for learning the duplicate detection technique. The more appropriate the property is selected, the more relevant duplicate can be identified.

RQ2: Which bug report properties are used for duplicate bug report detection?

The bug report properties can be also referred as bug report features where each feature indicates a new aspect of the reported bug. Based on the content type and previous usage, the features are divided into four major categories which are - Textual, Categorical, Contextual and Temporal.

5.1 Textual Feature

The textual feature refers to the bug report properties which are written in natural language format. The title, summary, description and comments are the main textual properties of a bug report as shown in Table 1. While reporting a bug, the developers generally add a short title/summary of the bug report which is written in natural language format. Besides the title, a detailed description of the bug is also added. The description often contains source code stack traces. As a result, the description of bug report may be of any length. After the reporting of the bug, developers interact with each other by posting comments during the bug resolution. Hence, the comment property also contains text in natural language format.

As the textual fields are written in natural language, so these fields indicate appropriate developer's or user's perspective about a reported bug. Table 1 shows that 20 out of the 23 papers have used description field of the bug report as feature for duplicate bug detection. Next, the title or summary is the most used bug report

feature. The table also shows that out of the four major categories, textual features are the most used bug report features.

Table 1: Features of Bug Report

Feature Type	Feature Name	No of Times Used
Textual	Description	20
	Title / Summary	18
	Comment	3
Categorical	Component	12
	Product	11
	Priority	10
	Operating System	7
	Version	5
	Severity	4
	Hardware	1
	Status	1
Contextual	Topic of report extracted based on textual features	7
Temporal	Open / Close Date	4
Identical	Bug Id	5
Structural / Attachment	Source code / file attachment	4

5.2 Categorical Feature

The categorical features represent the bug report properties which value is selected from a list of predefined values. It includes the component, product, priority, operating system, version, severity and status of the bug report. The similarity between two duplicate reports are checked by calculating the equality of the features. Table 1 shows that categorical features are the second highest used features by the existing techniques. The Product and Component are the most used features of categorical type.

5.3 Contextual Feature

Unlike the textual and categorical features, the contextual features tend to identify inherent topic of the bug report using topic modeling techniques (Rocha & Carvalho, 2021). These features are calculated by measuring the similarity between the content of the bug report and a predefined list of words corresponding to specific topics (such as security, performance, enhancement etc). The contextual similarity of bug reports are measured using different techniques such as Cosine similarity, Manhattan distance, LDA etc.

5.4 Temporal Feature

The temporal features tend to check the recency between the bug reports in terms of reporting or closing time. These features are calculated by taking the subtraction value of same fields from two different bug reports. The features are less used in comparison of the other features as shown in Table 1. However this feature can be helpful in filtering the recent bug reports. As a result, the search space of duplicate report checking can be reduced.

Apart from the above mentioned features, Identical, Structural and Derived features are also used in existing techniques (Mahfoodh & Hammad, 2022). The identical features check the distance between the unique id of bug reports to understand their reporting sequence. Source code patch or files are sometimes attached with the bug report (Wang et al., 2008). These file attachments are considered as structural features while duplicate detection. Derived features are calculated by applying the TF-IDF, BM25F and date interval calculation techniques on categorical and textual features of the bug report.

After extraction of features, ranking or combining features has also been done by few proposed techniques (Alkhazi et al., 2020, Neysiani et al., 2020). For ranking the features Naive aggregation, Ordinal regression and Learning-to-rank have been used in literature. Few existing works have assigned specific value to features for ranking. The values are assigned based on weighted average of information gain ratio, Gini index, chi-square, PCA of the features. The more appropriate feature is selected, the more appropriate duplicate reports can be identified. Therefore, techniques for ranking and identifying effective features can be explored further.

Since the bug report fields are stored in different formats as discussed above, therefore before checking similarity the data needs to be normalized in general format. Figure 4 shows the popular pre-processing techniques which have been used by researchers. It can be seen that text tokenization and stop word removal have been used 78.26% times in the 23 studied papers. Stemming is the third most used preprocessing technique being used in 43.48% cases. Jalbert et al. have applied MontyLingua tool, ReqSmile tool and Porter Stemming algorithm for tokenization, stop word removal and stemming respectively (Jalbert & Weimer, 2008). Lower case conversion, n gram word conversion, lemmatization are also used while pre-processing the text fields of the bug report. For stack trace or source code files, programming specific word removal, file path replacement mechanism are applied (Kumar et al., 2020).

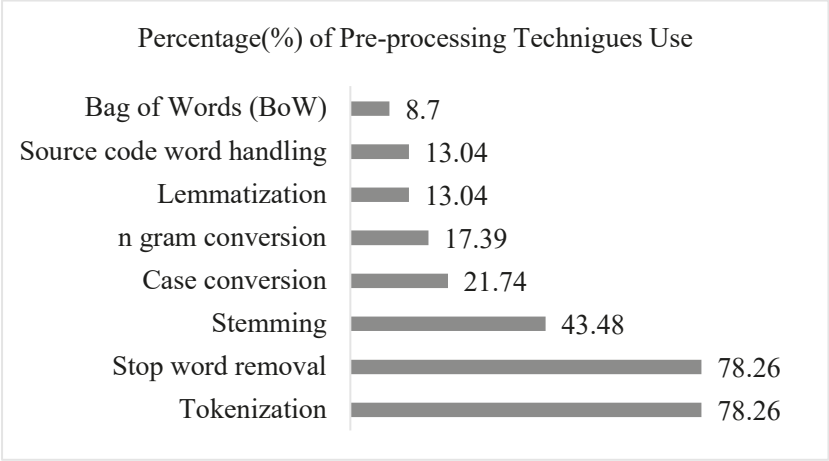


Figure 4: Sample eclipse bug report in XML format

Based on the above discussion, the answer to research question RQ2 depicts that textual and categorical properties are the most used properties of a bug report. On the other hand, stop word removal, tokenization and stemming are the popular text cleansing techniques. However, there is no clear discussion on among the four features, which are the most dominant one in identifying duplicates. Therefore, the effect of using individual and combined features in identifying duplicate bug reports can be studied in future.

6. Learning Algorithms

Once appropriate features are selected and pre-processed, the features are fed into different algorithms to develop a model that can detect duplicate bug reports.

RQ3: What type of algorithms are being used for duplicate bug report detection?

The existing techniques have used Machine Learning (ML) (Neysiani & Babamir, 2019c), Information Retrival (IR) (Sabor et al., 2017) and Deep Learning (DL) (Poddar et al., 2019) based techniques for identifying duplicate reports. Information retrieval based techniques generate a ranked list of duplicates corresponding to an incoming report (Sabor et al., 2017; Johannes & Mira, 2013) .On the other hand, machine learning and deep learning based techniques classify a incoming bug report as a duplicate (Akilan et al., 2020; Kukkar et al., 2020; He et al., 2020). Some of the existing work have also combined these techniques in different phases of the duplicate detection algorithm (Neysiani & Babamir, 2020).

For evaluation of the existing techniques, researchers have used different metrics such as Accuracy, Recall, Precision, F1- score etc. Among these, accuracy and

recall has been used by most of the techniques (Neysiani et al., 2020; He et al., 2020; Xiao et al., 2020; Neysiani & Babamir, 2020, 2019b). Accuracy refers how many correct classification have been predicted by the model as shown in Equation 1. On the other hand, recall refers how many correct duplicates have been placed in the ranking from the actual duplicates as shown in Equation 2.

$$Accuracy = \frac{True\ Prediction}{Total} \quad (1)$$

$$Recall = \frac{True\ Duplicates}{Actual\ Duplicates} \quad (2)$$

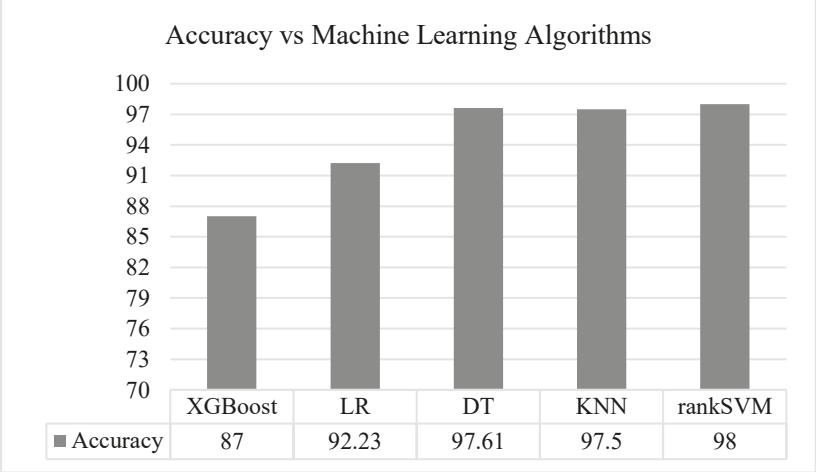


Figure 5: Analysis of accuracy in different Machine Learning (ML) algorithms

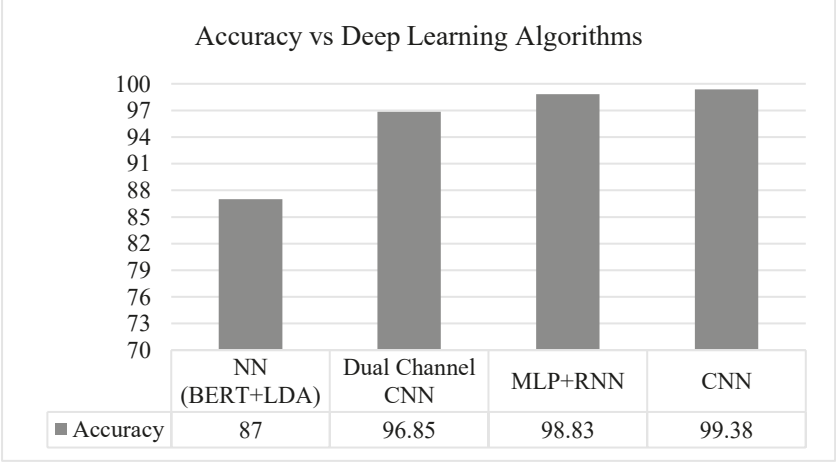


Figure 5 and 6 shows the accuracy of duplicate bug detection technique using different Machine Leaning and Deep Leaning algorithms respectively. The reason behind selecting these algorithms is the 23 reviewed papers have used these algorithms in combination or individual. Among the different ML algorithms the accuracy of Decision Tree (DT), KNN and rankSVM is above 97% which is prominent. On the other hand, the deep learning algorithms such as Recurrent Neural Network (RNN) and Convolutional Neural Network have accuracy above 98% which shows DL algorithms have better accuracy than ML algorithms. However, both of the algorithms have reached above 96% accuracy while detecting duplicate bug reports which indicate that there is small scope of future improvement in terms of accuracy. However, while achieving the accuracy the performance in terms of time, computation can be considered as future research scope. Besides the results of Figure 5 and 6 have been achieved using open source bug repositories where huge volume of data is available. The accuracy of these algorithms in terms of closed source bug repositories need further attention.

Table 2: Analysis of Recall in Information Retrieval, Machine Learning and Deep Learning Techniques

Information Retrieval Techniques			
LDA	TF-IDF	BM25F	LCS
67	84	93.04	97.27
Machine Learning Techniques			
LR	KNN	XGBoost	DT
92.95	97.51	98	99.94
Deep Learning Techniques			
NN	CNN	Dual Channel CNN	RNN
80	91.48	96.7	97.07

Table 2 shows the recall value of different Information Retrieval, Machine Learning and Deep Learning algorithms. Among the three category, IR based algorithms have lowest recall of 67% and 84% using Latent Dirichlet Allocation (LDA) and TF-IDF technique respectively. Jalber et al. also mentioned in the paper that TF-IDF achieved recall rate of 52% which is poor (Jalbert & Weimer, 2008). On the other hand, Machine Learning and Deep Learning techniques have higher recall in terms of IR based techniques. Decision Tree has accuracy of 99.94% in detecting duplicate bug reports (Soleimani & Morteza, 2020). The high value of recall for ML and DL techniques represent the limited scope of improvement in this metric. The future scope lies in evaluating the performance of these algorithms in terms of time, computation, memory usage etc. to reach this recall value.

The above discussion answers the research question RQ3 by identifying the fact that ML and DL algorithms achieve higher accuracy and recall than the IR based techniques. The higher value of these metrics show little scope of improvement in these metrics. However, in future the performance of combined application of these algorithm can be analyzed. Another future scope for research can be the implementation of duplicate detection technique as plugin for the software development IDEs. As a result, before reporting a bug, the developers can check for its duplicate bugs.

7. Conclusion

With the increasing of duplicate bug report submission, the need for appropriate duplicate bug report detection has become important. A general duplicate detection technique consists of three steps - *Feature Selection*, *Bug Report Pre-processing* and *Learning Algorithm or Model*. Based on this, a number of duplicate detection techniques have already been proposed by researchers. Therefore, this paper discusses the present literature work of duplicate bug report detection. In order to do so the papers devises three research questions which tend to analyze the input data set, the feature selection and prepossessing, and the evaluation of different learning algorithms respectively. With each research question, the future road map for research in duplicate detection has also been enlisted.

References

Akilan, T., Shah, D., Patel, N., & Mehta, R. (2020). Fast detection of duplicate bug reports using lda-based topic modeling and classification. In proceedings of the ieee international conference on systems, man, and cybernetics (smc) (pp. 1622–1629).

Alkhazi, B., DiStasi, A., Aljedaani, W., Alrubaye, H., Ye, X., & Mkaouer, M. W. (2020). Learning to rank developers for bug report assignment. Applied Soft Computing, 95 (pp. 106667).

Anvik, J., Hiew, L., & Murphy, G. C. (2005). Coping with an open bug repository. In proceedings of the oopsla workshop on eclipse technology exchange (pp. 35–39).

Bugzilla. (2021). Bugzilla addons. Retrieved from https://wiki.mozilla.org/Bugzilla:Addons (Last accessed 23 October 2021)

Cooper, N., Bernal-Cardenas, C., Chaparro, O., Moran, K., ´ & Poshyvanyk, D. (2021a). It takes two to tango: Combining visual and textual information for detecting duplicate video-based bug reports. In proceedings of the ieee/acm 43rd international conference on software engineering (icse) (pp. 957–969).

Cooper, N., Bernal-Cardenas, C., Chaparro, O., Moran, K., & ´ Poshyvanyk, D. (2021b). A replication package for it takes two to tango: Combining visual and textual information for detecting duplicate video-based bug reports. In proceedings of the ieee/acm 43rd international conference on software engineering: Companion proceedings (icse-companion) (pp. 160–161).

Ebrahimi, N., Trabelsi, A., Islam, M. S., Hamou-Lhadj, A., & Khanmohammadi, K. (2019). An hmm-based approach for automatic detection and classification of duplicate bug reports. Information and Software Technology, 113, (pp. 98–109).

Education, C.R.(2021).Core website. Retrieved from http://portal.core.edu.au/conf-ranks/ (Last accessed 19 November 2021)

He, J., Xu, L., Yan, M., Xia, X., & Lei, Y. (2020). Duplicate bug report detection using dual-channel convolutional neural networks. In proceedings of the 28th international conference on program comprehension (icpc) (pp. 117–127).

Jalbert, N., & Weimer, W. (2008). Automated duplicate detection for bug tracking systems. In proceedings of the ieee international conference on dependable systems and networks with ftcs and dcc (dsn) (pp. 52–61).

Jiang, Y., Lu, P., Su, X., & Wang, T. (2020). Ltrwes: A new framework for security bug report detection. Information and Software Technology, 124, (pp. 106314). Johannes, L., & Mira, M. (2013). Finding duplicates of your yet unwritten bug report. In proceedings of the 17th european conference on software maintenance and reengineering (pp. 69–78).

Kukkar, A., Mohana, R., Kumar, Y., Nayyar, A., Bilal, M., & Kwak, K.-S. (2020). Duplicate bug report detection and classification system based on deep learning technique. IEEE Access, 8, (pp. 200749–200763).

Kumar, A., Madanu, M., Prakash, H., Jonnavithula, L., & Aravilli, S. R. (2020). Advaita: Bug duplicity detection system. arXiv preprint arXiv:2001. (pp. 10376).

Lee, C.-P., & Lin, C.-J. (2014). Large-scale linear ranksvm. Neural computation, 26(4), (pp. 781–817).

Mahfoodh, H., & Hammad, M. (2022). Identifying duplicate bug records using word2vec prediction with software risk analysis. International Journal of Computing and Digital Systems, 11(1), (pp. 763–773).

Neysiani, B. S., & Babamir, S. M. (2019a). Duplicate detection models for bug reports of software triage systems: A survey. Current Trends In Computer Sciences & Applications, 1(5), (pp. 128–134).

Neysiani, B. S., & Babamir, S. M. (2019b). Improving performance of automatic duplicate bug reports detection using longest common sequence: Introducing new textual features for textual similarity detection. In proceedings of the 5th international conference on knowledge based engineering and innovation (kbei) (pp. 378–383).

Neysiani, B. S., & Babamir, S. M. (2019c). New methodology for contextual features usage in duplicate bug reports detection: dimension expansion based on manhattan distance similarity of topics. In proceedings of the 5th international conference on web research (icwr) (pp. 178–183).

Neysiani, B. S., & Babamir, S. M. (2020). Automatic duplicate bug report detection using information retrieval-based versus machine learning-based approaches. In proceedings of the 6th international conference on web research (icwr) (pp. 288–293).

Neysiani, B. S., Babamir, S. M., & Aritsugi, M. (2020). Efficient feature extraction model for validation performance improvement of duplicate bug report detection in software bug triage systems. Information and Software Technology, 126, (pp. 106344).

Poddar, L., Neves, L., Brendel, W., Marujo, L., Tulyakov, S., & Karuturi, P. (2019). Train one get one free: Partially supervised neural network for bug report duplicate detection and clustering. arXiv preprint arXiv:1903. (pp. 12431).

Rahman, M. M., Khomh, F., & Castelluccio, M. (2020). Why are some bugs non-reproducible?:–an empirical investigation using data fusion–. In proceedings of the 36th international conference on software maintenance and evolution (icsme) (pp. 605–616).

Rocha, T. M., & Carvalho, A. L. D. C. (2021). Siameseqat: A semantic context-based duplicate bug report detection using replicated cluster information. IEEE Access, 9, (pp. 44610–44630).

Sabor, K. K., Hamou-Lhadj, A., & Larsson, A. (2017). Durfex: a feature extraction technique for efficient detection of duplicate bug reports. In proceedings of the ieee international conference on software quality, reliability and security (qrs) (pp. 240–250).

Serrano, N., & Ciordia, I. (2005). Bugzilla, itracker, and other bug trackers. IEEE software, 22(2), (pp. 11–13).

Soleimani, N. B., & Morteza, B. S. (2020). Effect of typos correction on the validation performance of duplicate bug reports detection. In proceedings of the 10th international conference on information and knowledge technology (ikt), tehran, iran (pp. 1–2).

Tian, Y., Wijedasa, D., Lo, D., & Le Goues, C. (2016). Learning to rank for bug report assignee recommendation. In proceedings of the 24th international conference on program comprehension (icpc) (pp. 1–10).

Wang, X., Zhang, L., Xie, T., Anvik, J., & Sun, J. (2008). An approach to detecting duplicate bug reports using natural language and execution information. In proceedings of the 30th international conference on software engineering (pp. 461–470).

Xiao, G., Du, X., Sui, Y., & Yue, T. (2020). Hindbr: Heterogeneous information network based duplicate bug report prediction. In proceedings of the ieee 31st international symposium on software reliability engineering (issre) (pp. 195-206).

Zou, D., Deng, Z., Li, Z., & Jin, H. (2018). Automatically identifying security bug reports via multitype features analysis. In proceedings of the Australasian conference on information security and privacy (pp. 619–633)

Study of Spatio-temporal Variation of Humidity over the Southwestern Zone of Bangladesh

Md. Rakib Hasan[1] and Sirajul Hoque[2]

Abstract

In the last fifty years, the pattern of humidity has changed due to natural and anthropogenic reasons in the southwestern part of Bangladesh. The humidity data was recorded at eight regional meteorological stations of the Bangladesh Meteorological Department over the period of 1974 to 2020 and is used for assessments of trends of humidity aspects in the context of seasonal variability and spatial distribution in the southwestern zone of Bangladesh. For this study, the humidity trend was analyzed through Microsoft Excel Software, and the Arc GIS tool was used for spatial distribution analysis. In Khulna, during the dry season maximum mean humidity was78.6% in 2008 and the minimum mean humidity was 63% in 1974. In the wet season, maximum mean humidity reached 87.5% in 1984 and minimum mean humidity was 78.8% in 1976. In Jessore, the maximum mean humidity in the dry season was 78.8% in 1998 and the minimum mean humidity was 65% in 1975. In the wet season, the maximum mean humidity was 84% in 1990 and the minimum mean humidity was 75.1% in 1979. In Mongla during the dry season maximum mean humidity was 78.6% in 2005 and the minimum mean humidity was 69.6% in 1989. In 2005, during the wet season maximum mean humidity was 86.2% and the minimum mean humidity was 82.7% in 2014. In Satkhira during the dry season maximum mean humidity was 79.2% in 1998 and the minimum mean humidity was 58.4% in 1984. In the wet season, the maximum mean humidity was 82.8% in 1997 and the minimum mean humidity was 72.7 % in 1979. In Barishal during the dry season maximum mean humidity was 83.4% in 1986 and the minimum mean humidity was 71.6% in 1978. In the wet season, the maximum mean humidity was 88.4% in 1975 and the minimum mean humidity was 82.4% in 1982. In Bhola during the dry season maximum mean humidity was 83.6% in 1990 and the minimum mean humidity was 74.4% in 1978. In the wet season, the maximum mean humidity was 89.5% in 1991 and the minimum mean humidity was 82.5% in 2017. In Patuakhali during the dry season maximum mean humidity was 85.8% in 2010 and the minimum mean humidity was 65% in 1982. In 1979, during the rainy season maximum mean humidity was 90.2% and the minimum mean humidity was 78.8%. In Khepupara during the dry season maximum mean humidity was 89% in 1987 and the minimum mean humidity was

[1] Department of Environmental Science, Bangladesh University of Professionals, Dhaka, Bangladesh, Email: rakibhasanbup2021@gmail.com

[2] Department of Soil, Water and Environment, University of Dhaka, Dhaka, Bangladesh, Email: sirajswed@du.ac.bd

57% in 1979. In 1987, the rainy season's maximum mean humidity was 88% and the minimum mean humidity was 78.6%. in 1979. In the southwestern zone of Bangladesh, dry season humidity was consistently increasing trend while wet season humidity was decreasing.

Keywords: Spatial, Temporal, Relative Humidity, Variability, Trend Analysis, Climate, Interpolation.

1. Introduction

Climate change is a critical issue in today's modern world. A climate, according to the World Meteorological Organization, is defined as the 30-year average of meteorological parameters at a certain geographic place (Berger R, 2007). Various climatic variables, including rainfall, temperature, humidity, sunlight hour, and so on, have shown significant trends in different regions of the globe. Bangladesh is in the tropical monsoon area, with a climate defined by elevated temperatures, abundant rainfall, often extreme humidity, and significant seasonal fluctuations. It has dry and wet seasons. Due to climate change, Bangladesh is now one of the most climate-vulnerable states in the world, and it will become more so in the future. Floods, tropical cyclones, storm surges, and droughts are all expected to become more common and severe in the future years, according to climate scientists. In Bangladesh, the dry season is from November to March and the wet season is from April to October. During the dry and wet seasons, temperature, humidity, and rainfall change drastically. Temperature, precipitation, humidity, and wind speed fluctuations due to climate change have increased various problems throughout the world, including in Bangladesh. Khulna, Jessore, Satkhira, Barisal, Patuakhali, and Bhola is in the southwestern part of Bangladesh. Bangladesh is also known for its high humidity. When the water vapor content remains constant, whereas the temperature falls, the relative humidity rises. The relative humidity drops as the temperature rises, and the water vapor concentration remains constant. This is because colder air requires less moisture to become saturated than warmer air. The vapor has a significant impact on the dynamic properties of the global climate system. Humidity refers to the amount of water vapor in the air. In contrast, relative humidity encompasses the ratio of the current vapor pressure of the air to the saturated vapor pressure, which is usually stated as a percentage (Salman, 2020). Water vapor in the lower troposphere is the primary source of atmospheric precipitation and an essential component of the global water cycle. Moreover, water vapor is one of the greenhouse gases in the atmosphere, accounting for approximately 50% of the atmospheric greenhouse effect (Schmidt et al., 2010). Multiple climate models predict that the amount of water vapor in the atmosphere will rise because of global warming and that the additional water vapor

will exacerbate global warming due to the enhanced greenhouse effect, generating the increased destructive rebound. Bangladesh is a disaster-prone country. Over most of the western section of the country, March and April are the least humid months. In March, Dinajpur experienced the lowest average relative humidity (57%) of the year. January to March are the driest months in the eastern states. In March, Brahmanbaria recorded the lowest monthly average of 58.5 percent. From June to September, the relative humidity is somewhere over 80% everywhere. From 78.1 % in Cox's Bazar to 70.5 % in Pabna, the average relative humidity for the entire year is 78.1 % (Banglapedia, 2021). In this research, an attempt has been taken to analyze the spatial and temporal variation of humidity from 1974 to 2020 over a southwestern zone of Bangladesh.

2. Methodology

2.1 Study Area

In this study southwestern zone including Barishal, Bhola, Jessore, Khulna, Patuakhali, and Satkhira considered the study area. Barisal, Bhola, and Patuakhali districts belong to the Barisal division. Khulna, Satkhira and Jessore belongs to Khulna division. There are 19 coastal districts in Bangladesh, of which the above-mentioned districts are one of them. In these coastal districts, monsoon storms have a significant impact on agriculture. Figure1 illustrates the study area of the southwestern zone of Bangladesh.

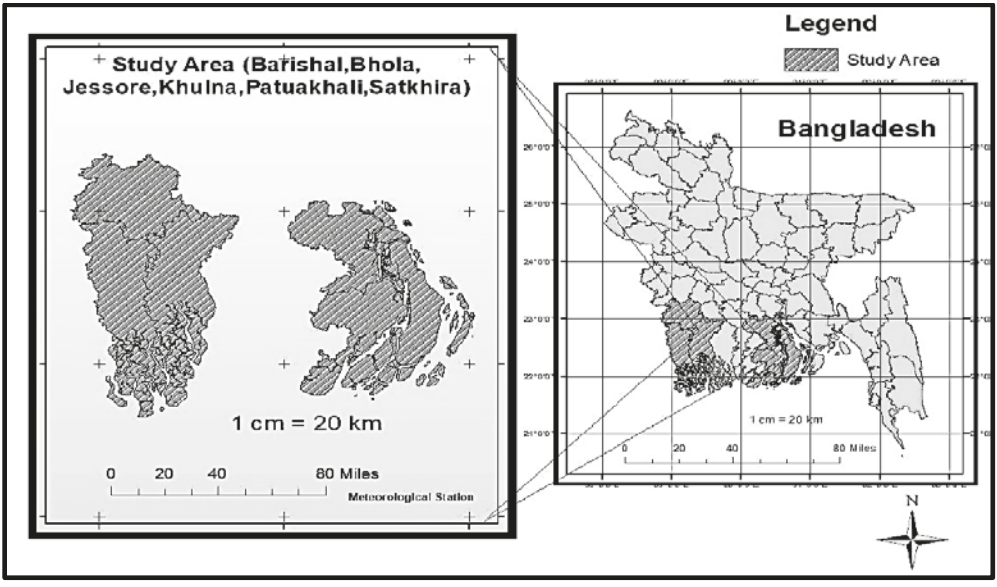


Figure 1: Study area of the south-western zone of Bangladesh

2.2 Geographic Location

Table-1: Geographic location of the study area

Station	Latitude	Longitude	Region
Barishal	22.717° N	90.367° E	Southwest
Bhola	22.683° N	90.650° E	Southwest
Jessore	23.184° N	89.161° E	West Central
Khepupara	21.983° N	90.2195° E	Southwest
Patuakhali	22.333° N	90.333° E	Southwest
Khulna	22.810° N	89.564° E	Southwest

2.3 Data Source and Methodology

The study used an empirical technique to investigate the temporal and spatial variation of humidity in Bangladesh's southwest and west-central regions over the last 46 years. The research utilized raw secondary data from 1974 to 2020 obtained from the Bangladesh Meteorological Department for 8 weather observation sites (BMD). The data has been analyzed by dividing the two seasons into 2 categories: dry and wet seasons. Microsoft Excel, python coding for time series analysis, and Arc GIS are used to analyze the spatial distribution map of the raw data. The majority of the graphs are bar charts, which are appropriate for demonstrating variation and trend analysis of mean humidity for the selected years 1974 to 2020. Regional variation in the humidity of BMD stations throughout the southern zone of Bangladesh has been illustrated using GIS software and the interpolation technique, which was implemented in this study.

$$m = \frac{sum\ of\ the\ terms}{number\ of\ terms} \quad \text{.....(1)}$$

Where,

m= mean

The following formula (equation -2) has been applied in to calculate the time series trend analysis of humidity from 1974 to 2020 over the southwestern and west-central zone of Bangladesh during the dry and wet season

$$y=bx+a \quad \text{.....(2)}$$

Where,

y - the dependent variable

x - the independent variable

a - the intercept

b - the slope

The following formula (equation -3) has been applied in to calculate the standard deviation of temperature, rainfall, and humidity from 1974 to 2020 over the south-western and west-central zone of Bangladesh during the dry and wet season

$$\sigma = \sqrt{\frac{\sum(xi - \mu)^2}{N}}$$

Where,

$\sigma =$ population standard deviation,

$N =$ the size of the population

$xi =$ each value from the population

$\mu =$ the population mean

3. Result and Discussion

3.1 Variation of Humidity Pattern over the southwestern zone of Bangladesh 1974 -2020

Figure 2 illustrates the seasonal variation of humidity in Khulna (1974-2020), the dry season. For this investigation, data on the relative humidity in Khulna is collected from Bangladesh Meteorological Department from 1974 to 2020. Bangladesh experienced significantly lower humidity during the dry season than in the wet season. Figure 2 shows that the relative humidity at Khulna in the dry season has quite an increasing trend. In the dry season, maximum mean humidity was observed at 78.6% in 2008 and minimum mean humidity was observed at 63% in 1974. In the dry season, the highest rate of humidity was found in January, November, and December. Humidity was above 80% in the mentioned months. There are two potential causes for the rising trend in humidity: one is related to increasing temperatures, and the other is due to increased land surface wetness.

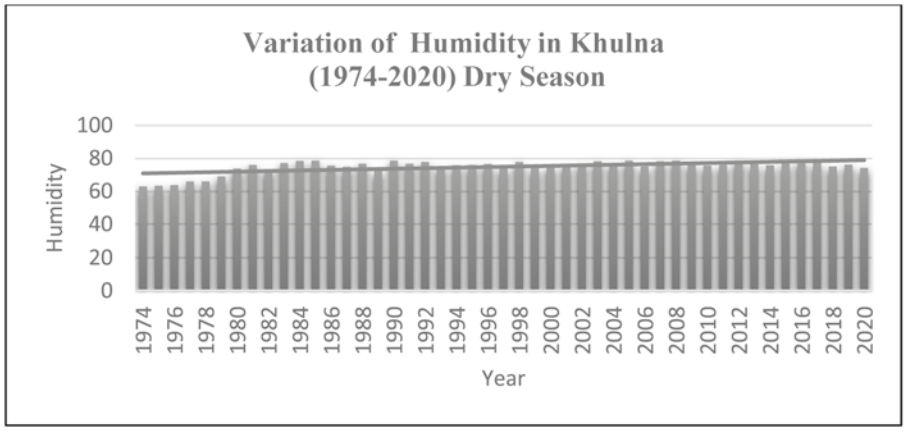


Figure 2: Seasonal variation of humidity in Khulna (1974-2020), dry season

Figure 3 illustrates the seasonal variation of humidity in Khulna (1974-2020), the wet season. For this investigation, data on the relative humidity in Khulna is collected from Bangladesh Meteorological Department from 1974 to 2020. During the dry season, Bangladesh experienced much lower humidity than in the wet season. Graph 3 shows that the relative humidity at Khulna in the wet season has quite a decreasing trend. In the wet season, maximum mean humidity was observed at 87.5 % in 1984 and minimum mean humidity was observed at 78.8 % in 1976. Based on historical statistics, the humidity level in Khulna is being substantially higher essentially every month during the wet seasons

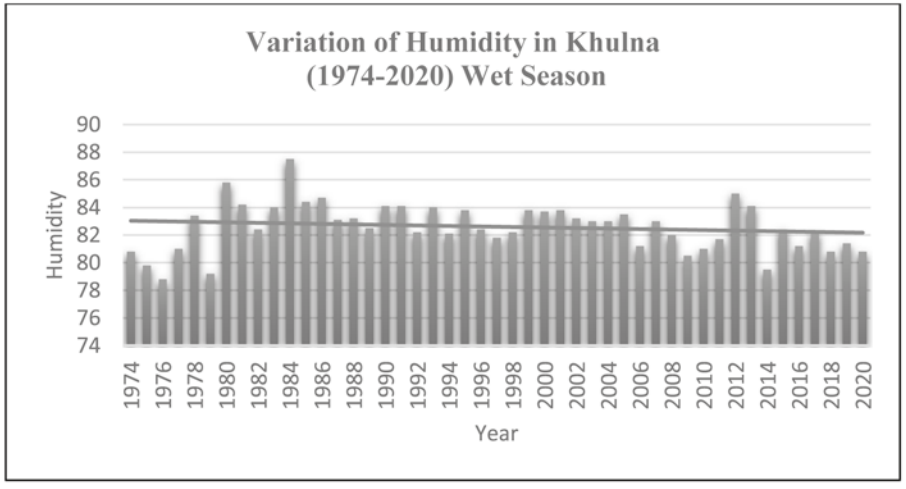


Figure 3: Seasonal variation of humidity in Khulna (1974-2020), wet season

Figure 4 depicts the seasonal variation of humidity in Jessore (1974-2020), the dry season. For this investigation, data on the relative humidity in Jessore is collected from Bangladesh Meteorological Department from 1974 to 2020. The dry season in Bangladesh has been much cooler than the wet season. Figure 4 shows that the relative humidity at Jessore in the dry season has quite an increasing trend. In the dry season, maximum mean humidity was observed at 78.8 % in 1998 and minimum mean humidity was observed at 65 % in 1975. Humidity has been on the increase over the past 10 years, from 1989 to 1999.

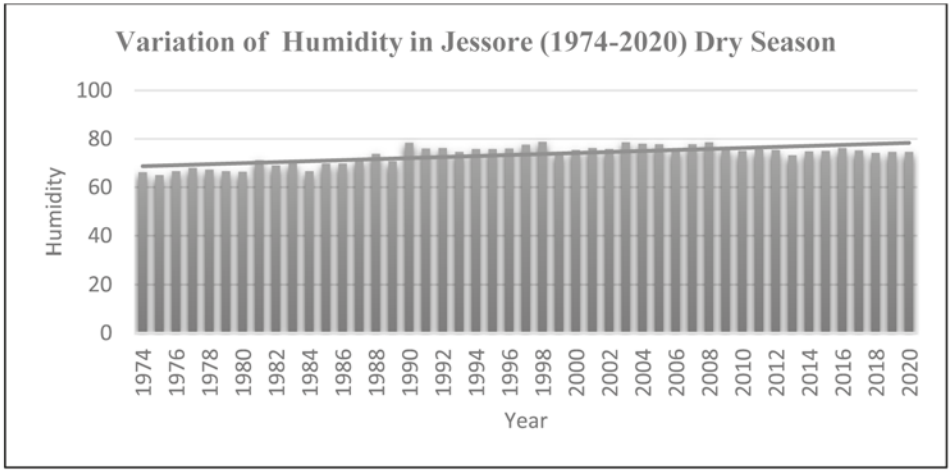


Figure 4: Seasonal variation of temperature in Jessore (1974-2020), dry season

Figure 5 depicts the seasonal variation of humidity in Jessore (1974-2020), the wet season. For this investigation, data on the relative humidity in Jessore is collected from Bangladesh Meteorological Department from 1974 to 2020. The dry season in Bangladesh often had less humidity than that of the wet season. Figure 5 shows that the relative humidity at Jessore in the wet season has quite a decreasing trend. In the wet season, maximum mean humidity was observed at 84 % in 1990 and minimum mean humidity was observed at 75.1 % in 1979. Humidity has been on the increase significantly since 1990 and 1993.

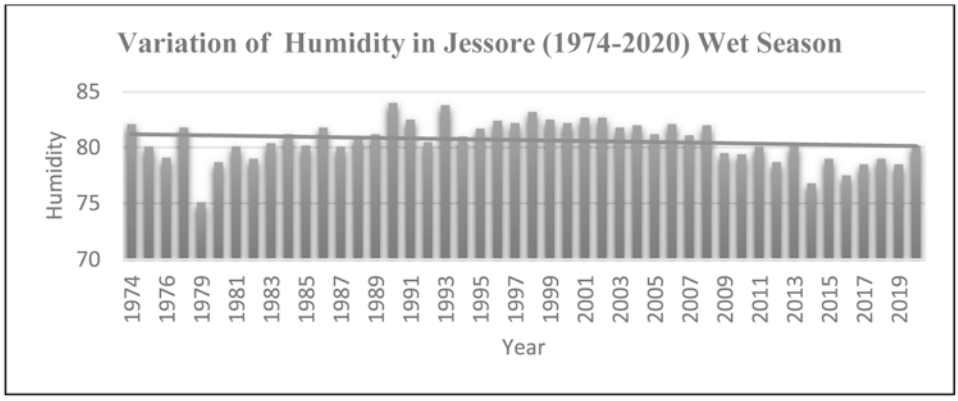


Figure 5: Seasonal variation of temperature in Jessore (1974-2020), wet season

Figure 6 depicts the seasonal variation of humidity in Mongla (1989-2020), during the dry season. For this investigation, data on the relative humidity in Jessore is collected from Bangladesh Meteorological Department from 1989 to 2020. Even during the dry season, Bangladesh received considerably low humidity than it was during the wet season. Figure 6 shows that the relative humidity at Mongla in the dry season has quite an increasing trend. In the dry season, maximum mean humidity was observed at 78.6 % in 2005 and minimum mean humidity was observed at 69.6 % in 1989. Between 1990 and 1993, the humidity level increased. In the following years, from 2003 to 2005 and 2009 to 2013 there is also a variety of humidity.

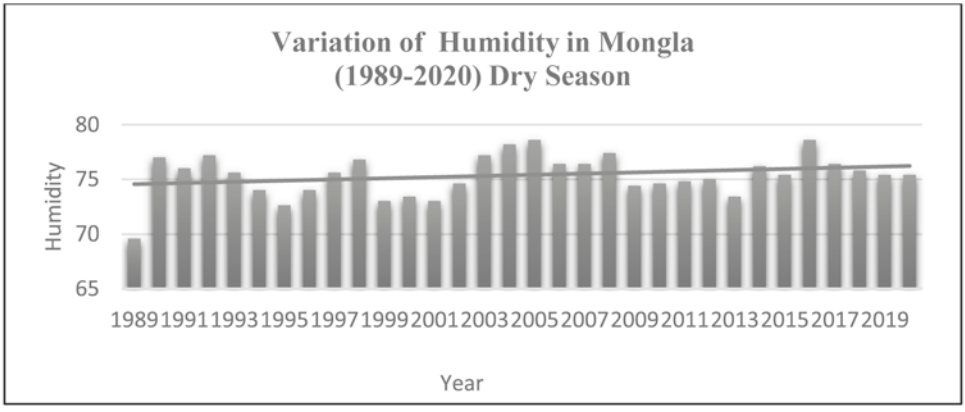


Figure 6: Seasonal variation of humidity in Mongla (1989-2020), dry season

Figure 7 illustrates the seasonal variation of humidity in Mongla (1989-2020), during the wet season. For this investigation, data on the relative humidity in Jessore is collected from Bangladesh Meteorological Department from 1974 to

2020. Bangladesh experienced significantly lower humidity during the dry season than in the wet season. Figure 7 shows that the relative humidity at Mongla in the wet season has quite a decreasing trend. In the wet season, maximum mean humidity was observed at 86.2 % in 2005 and minimum mean humidity was observed at 82.7 % in 2014. Humidity has been on the increase significantly from 2004 to 2006. From 2008 to 2019, humidity is rather low, but it began to rise again in 2020.

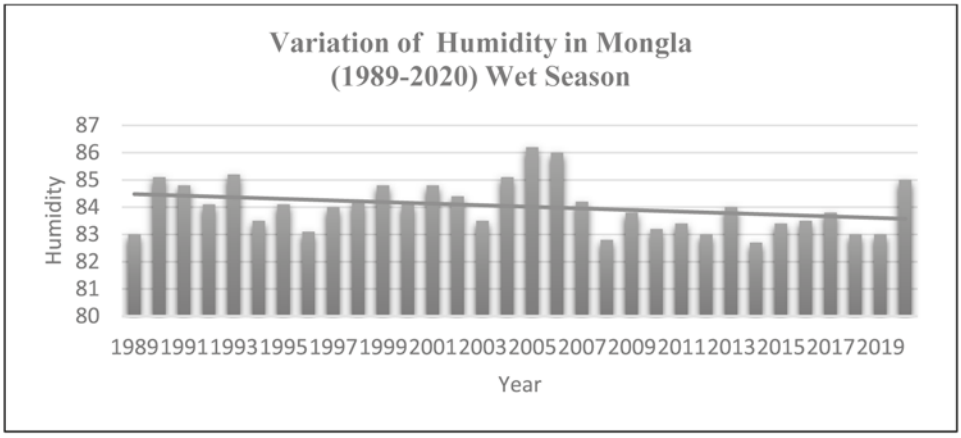


Figure 7: Seasonal variation of humidity in Mongla (1989-2020), wet season

Figure 8 depicts the seasonal variation of humidity in Satkhira (1974-2020), the dry season. For this investigation, data on the relative humidity in Jessore is collected from Bangladesh Meteorological Department from 1974 to 2020. Bangladesh experienced significantly lower humidity during the dry season than in the wet season. Figure 8 shows that the relative humidity at Satkhira in the dry season has quite an increasing trend. In the dry season, maximum mean humidity was observed at 79.2 % in 1998 and minimum mean humidity was observed at 58.4 % in 1984. In the following years, from 1982 to1988 there was also a variety of humidity.

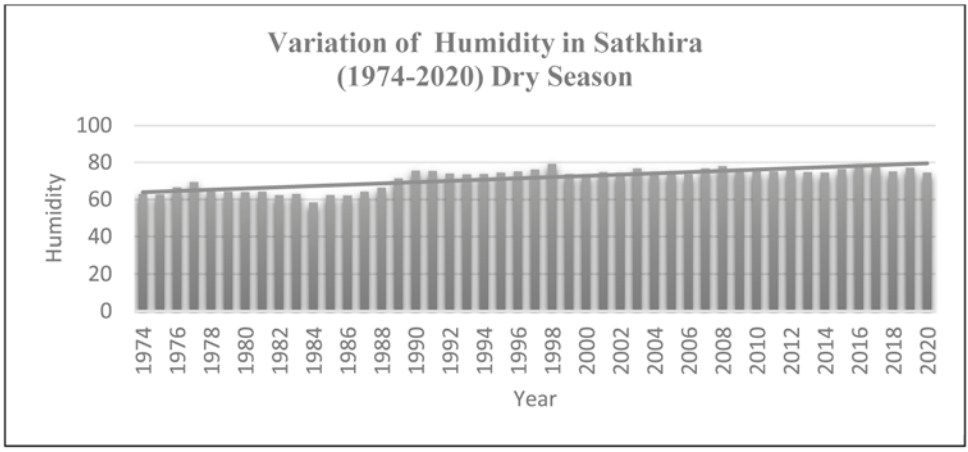


Figure 8: Seasonal variation of humidity in Satkhira (1974-2020), dry season

Figure 9 illustrates the seasonal variation of humidity in Satkhira (1974-2020), during the wet season. For this investigation, data on the relative humidity in Jessore is collected from Bangladesh Meteorological Department from 1974 to 2020. Figure 9 shows that the relative humidity at Satkhira in the wet season has quite an increasing trend. In the wet season, maximum mean humidity was observed at 82.8 % in 1997 and minimum mean humidity was observed at 72.7 % in 1979. Humidity has been on the decrease from 1979,1982 and 1984 to 1986. Humidity has been on the increase significantly from 1989 to 1992 and 1997 to 1999.

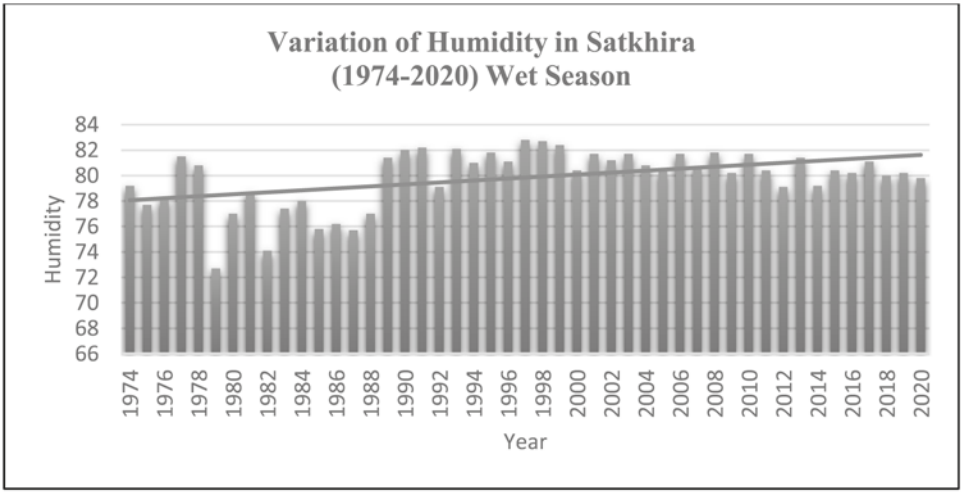


Figure 9: Seasonal variation of humidity in Satkhira (1974-2020), wet season

Figure 10 depicts the seasonal variation of humidity in Barishal (1974-2020), the dry season. For this investigation, data on the relative humidity in Barishal is collected from Bangladesh Meteorological Department from 1974 to 2020. Figure 10 shows that the relative humidity at Barishal in the dry season has quite an increasing trend.

In the dry season, maximum mean humidity was observed at 83.4 % in 1986 and minimum mean humidity was observed at 71.6 % in 1978. In the following years, from 1982 to1988, 1990 to 2005 there was also a variety of humidity.

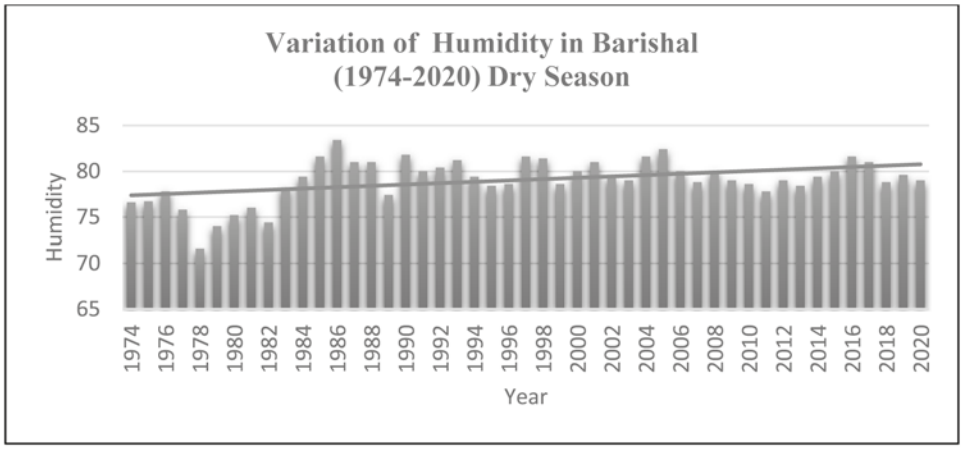


Figure 10: Seasonal variation of humidity in Barishal (1974-2020), dry season

Figure 11 illustrates the seasonal variation of humidity in Barishal (1974-2020), during the wet season. For this investigation, data on the relative humidity in Barishal is collected from Bangladesh Meteorological Department from 1974 to 2020. Figure 11 shows that the relative humidity at Barishal in the wet season has quite a decreasing trend. In the wet season, maximum mean humidity was observed at 88.4 % in 1975 and minimum mean humidity was observed at 82.4 % in 1982. Humidity has been on the decrease since 1976,1979,1982 and 2014. Humidity has been on the increase significantly from 1975, 1985,1993,2003, and 2017 .

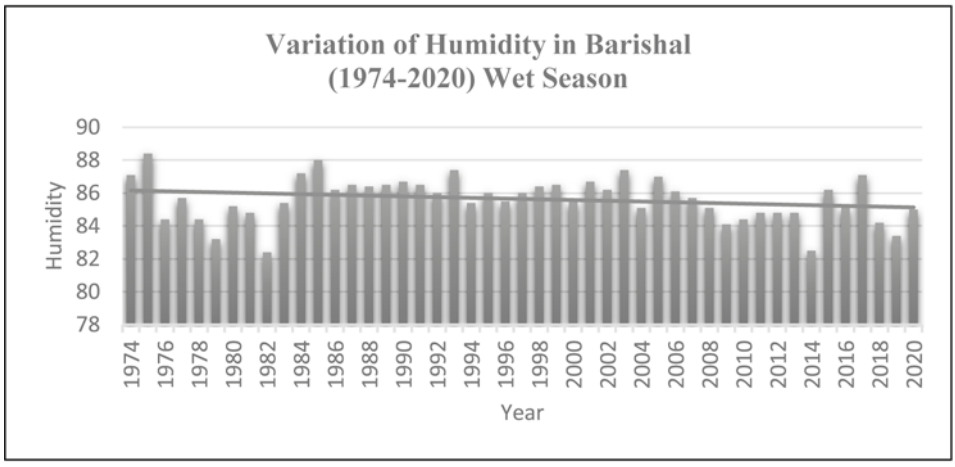


Figure 11: Seasonal variation of humidity in Barishal (1974-2020), wet season

Figure 12 depicts the seasonal variation of humidity in Bhola (1974-2020), the dry season. For this investigation, data on the relative humidity in Bhola is collected from Bangladesh Meteorological Department from 1974 to 2020. Figure 12 shows that the relative humidity at Bhola in the dry season has quite an increasing trend. In the dry season, maximum mean humidity was observed at 83.6 % in 1990 and minimum mean humidity was observed at 74.4 % in 1978. In the following years, from 1982 to1988 humidity has been increased significantly.

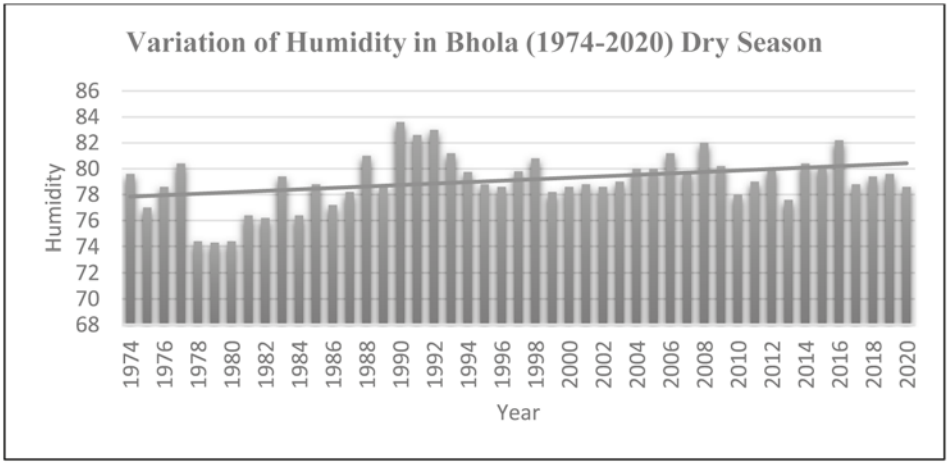


Figure 12: Seasonal variation of humidity in Bhola (1974-2020), dry season

Figure 13 illustrates the seasonal variation of humidity in Bhola (1974-2020), during the wet season. For this investigation, data on the relative humidity in Bhola

is collected from Bangladesh Meteorological Department from 1974 to 2020. Figure 13 shows that the relative humidity at Bhola in the wet season has quite a decreasing trend.

In the wet season, maximum mean humidity was observed at 89.5 % in 1991 and minimum mean humidity was observed at 82 % in 2017. Humidity has been on the decrease since 1982,1985,1986,2014, and 2017. Humidity has been on the increase significantly between 1974 and 1991.

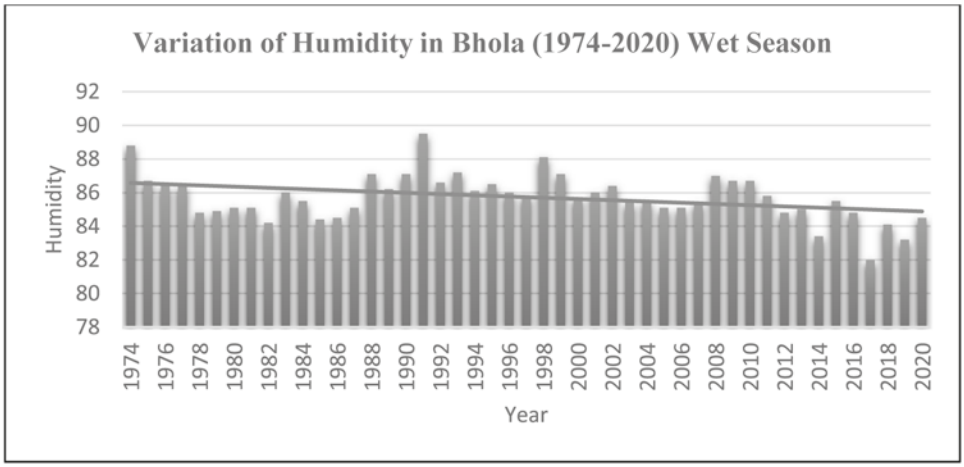


Figure 13: Seasonal variation of humidity in Bhola (1974-2020), wet season

Figure 14 depicts the seasonal variation of humidity in Patuakhali (1975-2020), the dry season. For this investigation, data on the relative humidity in Patuakhali is collected from Bangladesh Meteorological Department from 1975 to 2020. Figure 14 shows that the relative humidity at Patuakhali in the dry season has quite an increasing trend. In the dry season, maximum mean humidity was observed at 85.8 % in 2010 and minimum mean humidity was observed at 65 % in 1982. In the following years, between 2003 and 2010 humidity increased significantly.

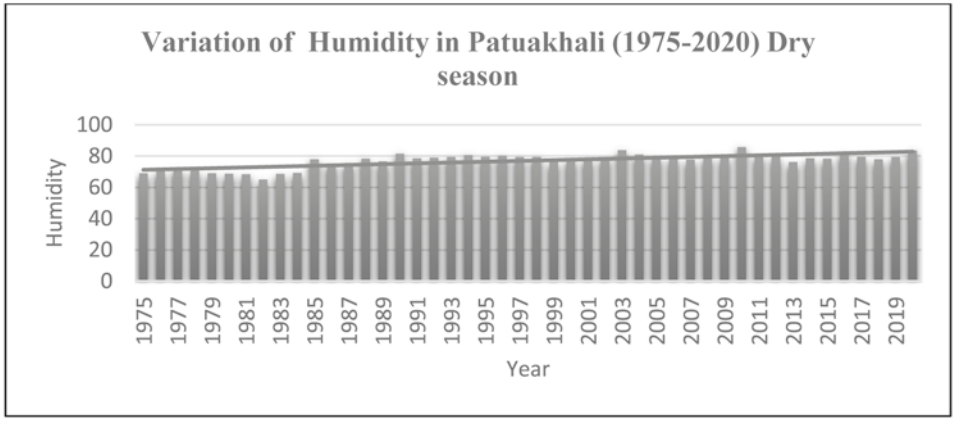


Figure 14: Seasonal variation of humidity in Patuakhali (1975-2020), dry season
Figure 15 illustrates the seasonal variation of humidity in Patuakhali (1974-2020), during the wet season. For this investigation, data on the relative humidity in Patuakhali is collected from Bangladesh Meteorological Department from 1974 to 2020. Figure 15 shows that the relative humidity at Patuakhali in the wet season has quite an increasing trend. In the wet season, maximum mean humidity was observed at 90.2 % in 2010 and minimum mean humidity was observed at 78.8 % in 1979. Humidity has been on the decrease since 1980 and 1982. Humidity has been on the increase significantly from 2001 to 2005, 2010, and 2020.

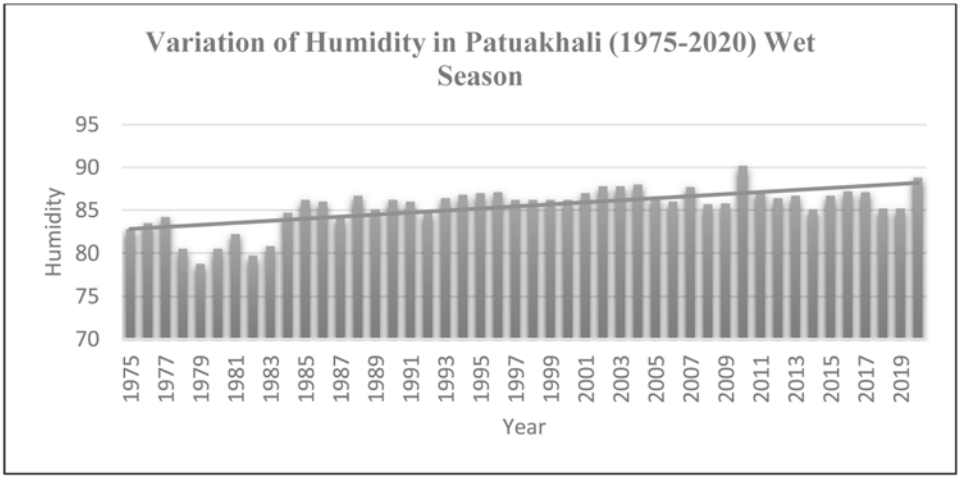


Figure 15: Seasonal variation of humidity in Patuakhali (1975-2020), wet season
Figure 16 depicts the seasonal variation of humidity in Khepupara (1974-2020), the dry season. For this investigation, data on the relative humidity in Khepupara is collected from Bangladesh Meteorological Department from 1975 to 2020.

Figure 16 shows that the relative humidity at Khepupara in the dry season has quite an increasing trend.

In the dry season, maximum mean humidity was observed at 89 % in 1987 and minimum mean humidity was observed at 57 % in 1979. In the following years, between 1985 and 1988 humidity increased significantly.

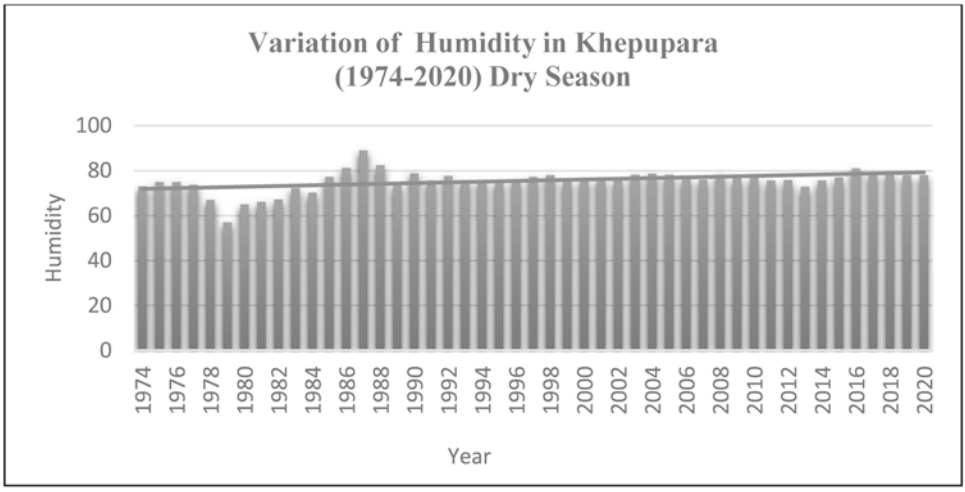


Figure 16: Seasonal variation of humidity in Khepupara (1974-2020), dry season

Figure 17 illustrates the seasonal variation of humidity in Khepupara (1974-2020), during the wet season. For this investigation, data on the relative humidity in Khepupara is collected from Bangladesh Meteorological Department from 1974 to 2020. Figure 17 shows that the relative humidity at Khepupara in the wet season has quite an increasing trend. In the wet season, maximum mean humidity was observed at 88 % in 1987 and minimum mean humidity was observed at 78.6 % in 1979. Humidity has been on the decrease from 1978 to 1980 and 1982 to 1985. Humidity has been on the increase significantly from 1986 to 1987 and in 2013.

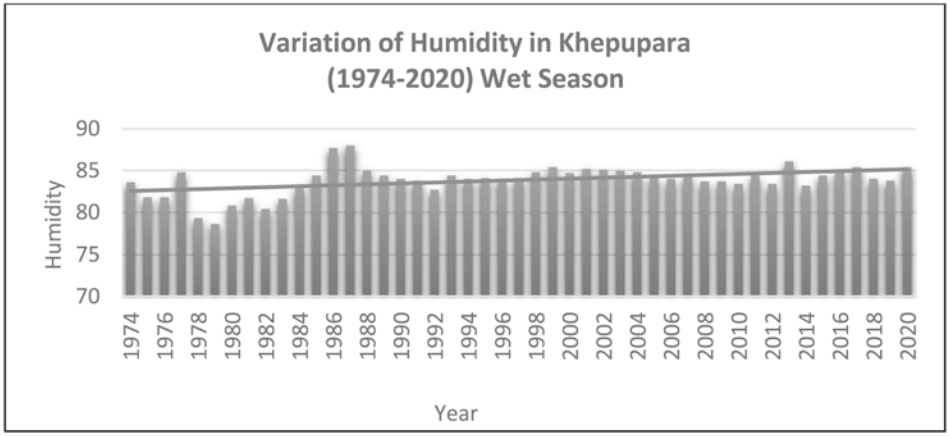


Figure 17: Seasonal variation of humidity in Khepupara (1974-2020), wet season

3.2 Spatial Distribution of Humidity over the Southwestern Zone of Bangladesh 1974 -2020

Figure 18 illustrates the spatial distribution of the mean annual humidity of the dry season (1974-2020) using the technique of inverse distance weighted (IDW) interpolation.

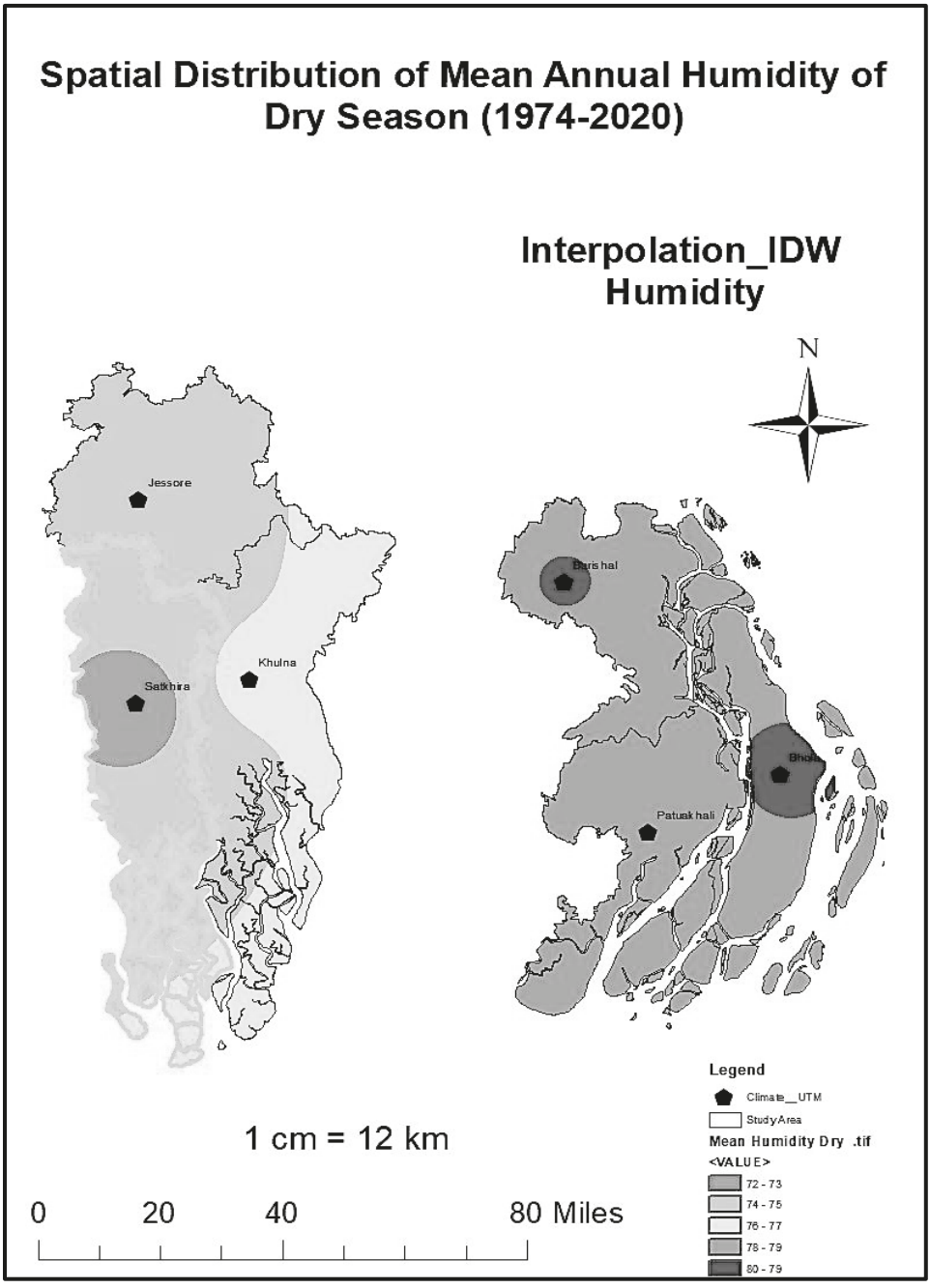


Figure 18: Spatial distribution of mean annual humidity of dry season (1974-2020) by using inverse distance weighted technique

Figure 19 illustrates the spatial distribution of the mean annual humidity of the dry season (1974-2020) using the kriging interpolation technique.

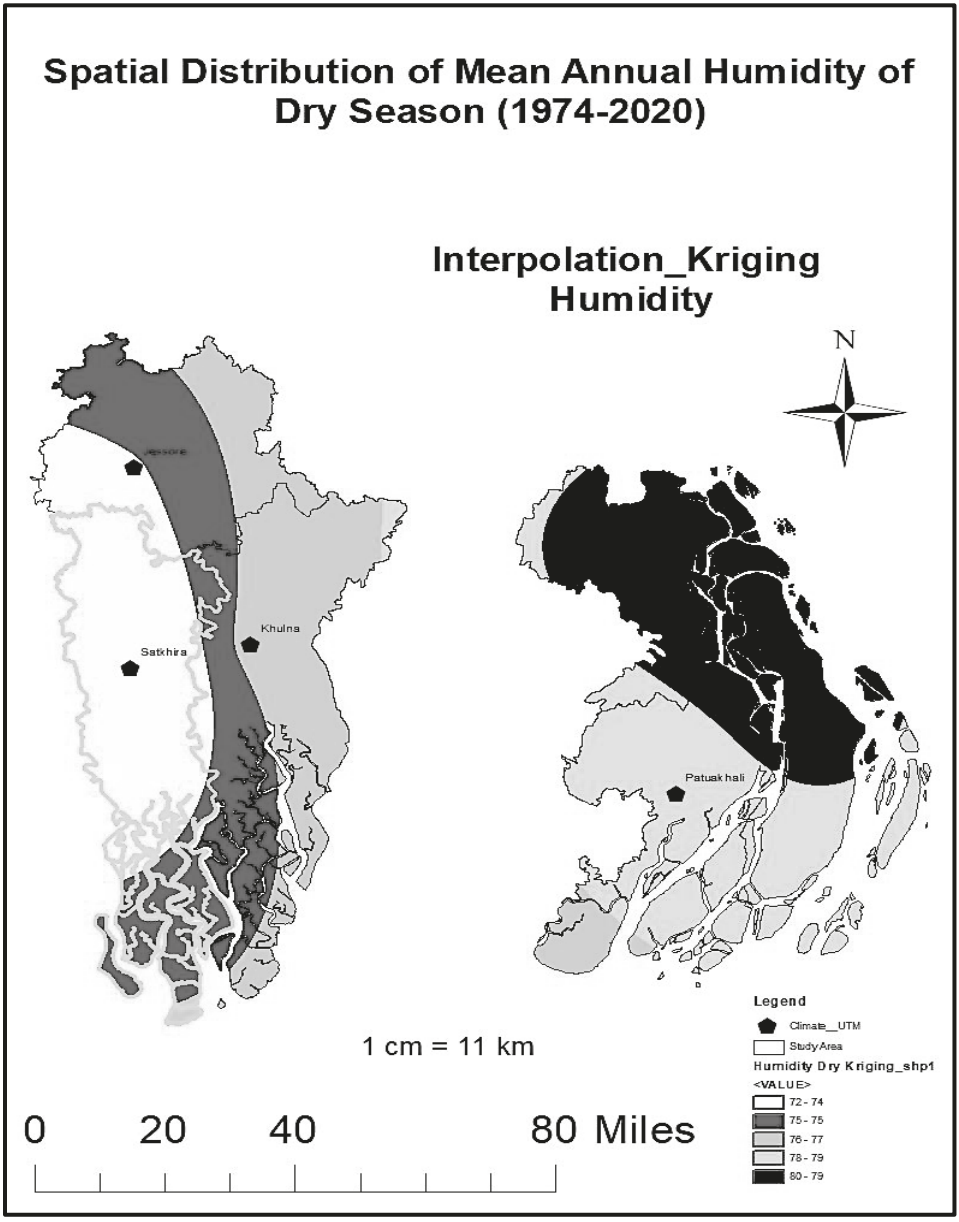


Figure 19: Spatial distribution of mean annual humidity of dry season (1974-2020) by using the kriging technique

The spatial distribution map of figures 18 and 19 shows that in the dry season highest humidity is 79 % at Bhola and the lowest at 71% at Satkhira.

Figure 20 illustrates the spatial distribution of the mean annual humidity of the wet season (1974-2020) using the technique of inverse distance weighted (IDW) interpolation.

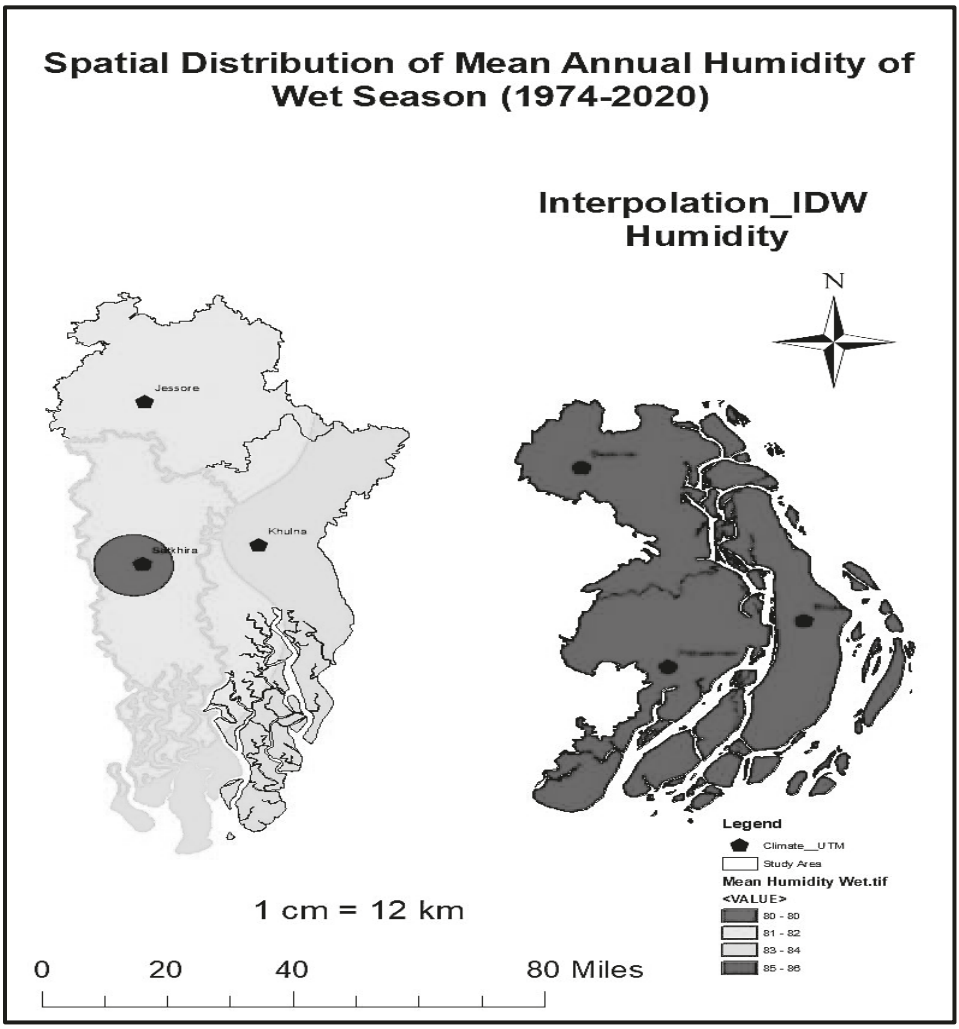


Figure 20: Spatial distribution of mean annual humidity of wet season (1974-2020) by using inverse distance weighted technique

Figure 21: illustrates the spatial distribution of the mean annual humidity of the wet season (1974-2020) using the Kriging interpolation technique.

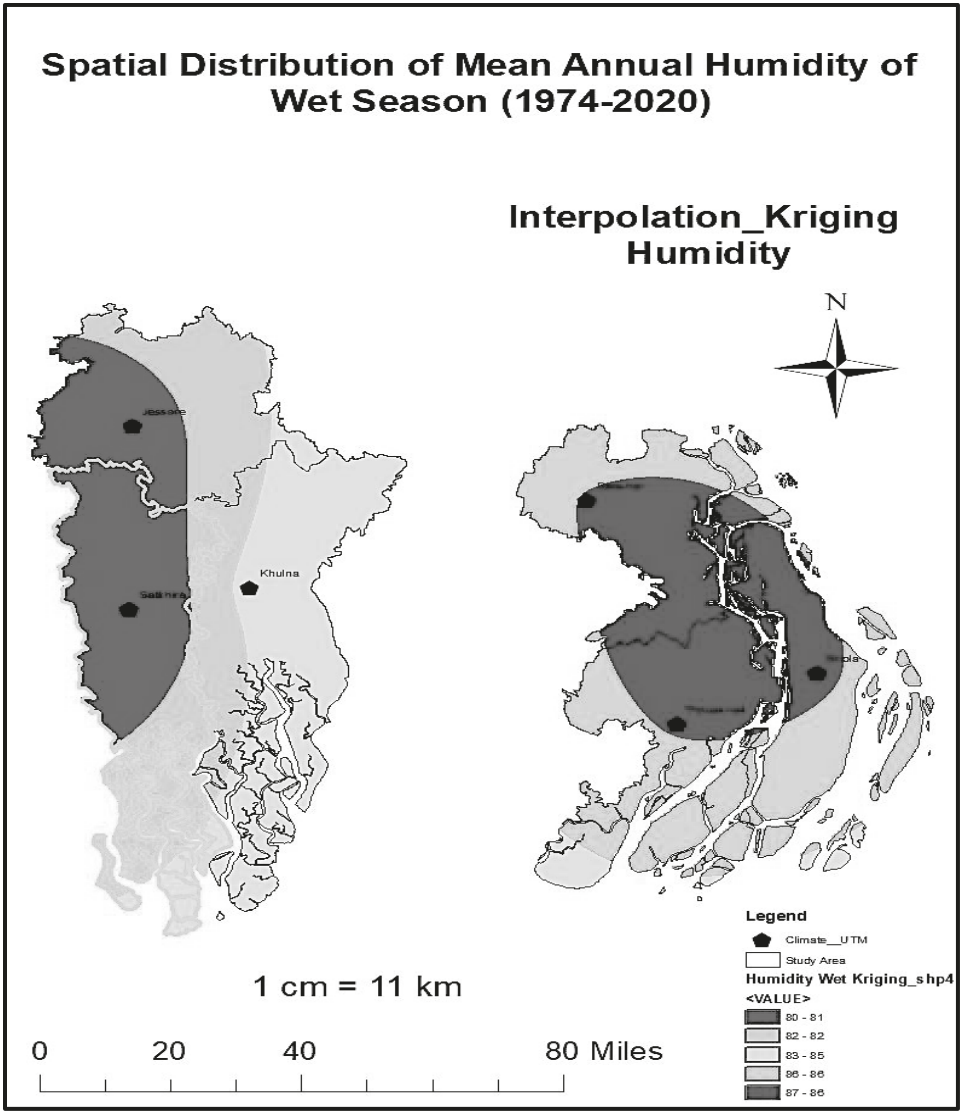


Figure 21: Spatial distribution of mean annual humidity of wet season (1974-2020) by using the kriging technique

The spatial distribution map of figures 20 and 21 shows that in the wet season highest humidity is 85 % at Bhola and the lowest at 79% at Satkhira.

3.3 Comparative Variation of Humidity in the Southwestern Zone of Bangladesh

Figure 22 illustrates the comparative analysis of mean annual humidity in the southwestern zone of Bangladesh wherein the wet season Bhola area has the maximum mean humidity at 85.7% and mean minimum humidity of 79.8 % in the Satkhira area. In the dry season, the Bhola area has the maximum mean humidity of 79.2 % and the minimum mean humidity of 71.7% in the Satkhira area.

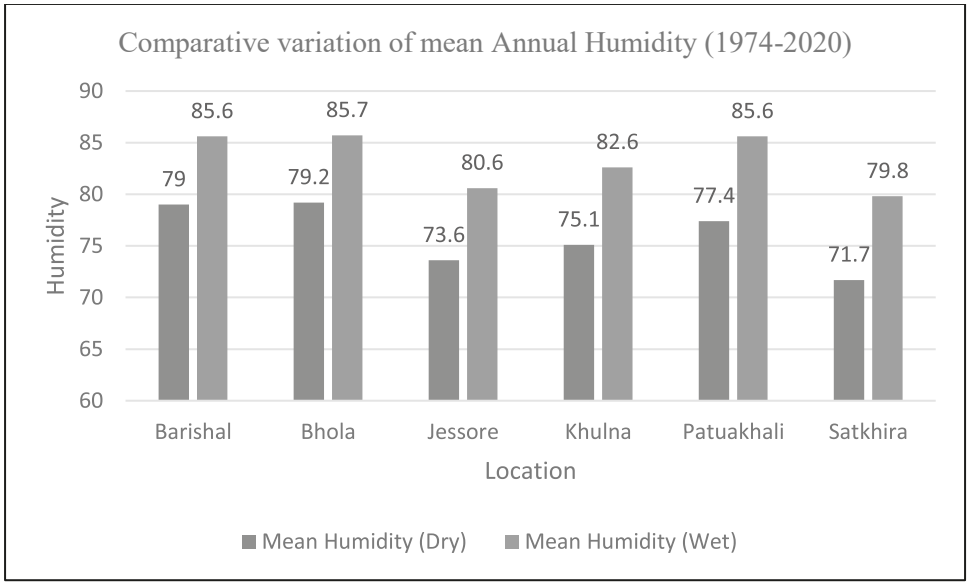


Figure 22: Comparative analysis of mean annual humidity in the southwestern zone of Bangladesh

4. Conclusion

This study assessed the spatial and temporal variation of humidity over the southwestern zone of Bangladesh from 1974 to 2020. Analysis of dry and wet season data of the last 50 years shows that dry season humidity was always upward while wet season humidity was downward in the southwestern region of Bangladesh. However, especially in the Patuakhali region, the trend of humidity

was on the rise during the wet season. The average monthly relative humidity ranges from 45 % in March to 79 % in June, with an average annual relative humidity of 65.8%. The ideal humidity is 30 to 60 %. However, an analysis of the data from the last 50 years in the southwestern region of Bangladesh indicates that humidity has always been above 70% or more in the dry season and above 80% or more in the wet season. Due to climate change, there has been a huge fluctuation in humidity in the southwestern region of Bangladesh.

References

Banglapedia,2021, Available at: https://en.banglapedia.org/index.php/Climate (Accessed on 19-06-2021).

Berger, R., 2007, February. Adapting to drought in Kenya: The experience of Practical Action in central Kenya. In 2nd International Workshop on Community Based Adaptation to Climate Change. Organized by: Bangladesh Centre for Advanced Studies, IIED & The Ring. Radisson Water Garden Hotel, Dhaka, 24th–28th February.

Musashi, J.P., Pramoedyo, H. and Fitriani, R., 2018. Comparison of inverse distance weighted and natural neighbor interpolation method at air temperature data in Malang region. Cauchy, 5(2), pp.48-54.

Philip, G.M. and Watson, D.F., 1982. A precise method for determining contoured surfaces. The APPEA Journal, 22(1), pp.205-212.

Salman, M.A. and Ahmed, F., 2020. Climatology in Barishal, Bangladesh: A Historical Analysis of Temperature, Rainfall, Wind Speed and Relative Humidity Data. Malaysian Journal of Geosciences (MJG), 4(1), pp.43-53.

Schmidt, G.A., Ruedy, R.A., Miller, R.L. and Lacis, A.A., 2010. Attribution of the present-day total greenhouse effect. Journal of Geophysical Research: Atmospheres, 115(D20). Available at: http://dx.doi.org/10.1029/2010JD014287

Investigation of Depression Using Context Analysis

Salma Akter Asma[1], Sadik Hasan[2], Nazneen Akhter[3], Mehenaz Afrin[4], Afrina Khatun[5], Kazi Abu Taher[6]

Abstract

Depression is a major concern in today's time as it is becoming a pandemic worldwide. Nowadays people (especially the young generation) are using social media sites to share their feelings, emotions, and personal life activities. Their mental health condition can be analysed by reviewing their social media posts and activities. Recent research work in this field is trying to go beyond manual depression detection. Hence, an automated system is necessary for analysing depression symptoms from social media for the sake of society. For this purpose, in this work, a Machine Learning based depression detection technique has been proposed. To develop the model six Machine Learning (ML) classifiers namely Support Vector Machine (SVM), Decision Tree (DT), K-Nearest Neighbour (KNN), Passive Aggressive (PA), Random Forest (RF), and Bagging classifier have been used. To improve the performance of the classifiers a dimension reduction technique namely Latent Semantic Analysis (LSA) is used. A comparison among four-dimension reduction techniques such as Latent Semantic Analysis (LSA), Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Fast Independent Component Analysis (Fast ICA) is given to justify why LSA is considered a dimension reduction technique in this work. With LSA, the Bagging classifier reached the top performance with an accuracy of 94.62%, while the base classifier is RF.

Keywords: Machine Learning, Depression, Dimension Reduction, Contextual Meaning, LSA.

---

[1] Department of Information and Communication Technology, Bangladesh University of Professionals, Dhaka, Bangladesh, Email: aaliyaasma786@gmail.com

[2] Department of Information and Communication Technology, Bangladesh University of Professionals, Dhaka, Bangladesh, Email: sadik.h.emon@gmail.com

[3] Department of Computer Science and Engineering, Bangladesh University of Professionals, Dhaka, Bangladesh, Email: nazneen.akhter@bup.edu.bd

[4] Department of Information and Communication Technology, Bangladesh University of Professionals, Dhaka, Bangladesh, Email: agroti202@gmail.com

[5] Department of Information and Communication Technology, Bangladesh University of Professionals, Dhaka, Bangladesh, Email: afrina.khatun@bup.edu.bd

[6] Department of Information and Communication Technology, Bangladesh University of Professionals, Dhaka, Bangladesh, Email: kataher@bup.edu.bd

1. Introduction

Nowadays, depression is the most alarming disorder that has spread mental illness globally. More than 264 million people of different ages have been suffering from this illness (Safiri et al., 2022).

Though depression often goes unnoticed by people for the maximum time, the consequences can be devastating. According to the World Health Organization (WHO) survey in 2012, nearly one million people commit suicide yearly due to depression (Hassan et al., 2017). They estimate that around 322 million people worldwide will have mental illness within 2030 (Hur et al., 2018). However, approximately 70% of patients would not consult a doctor in the primary state of depression which is the manifestation of extreme unconsciousness (Murray & Lopez, 1997). The governance of mental health issues is so different because of the dynamic variety of human psychology, complex to fetch the pattern of mental behavior, mental illness is noticed in the last stage, and people normally refuse it fearing culture and stigma (Gupta et al., 2021).

At present, more than half of the world's population is using social media sites. The trend of using social media has increased, especially in pandemic situations. By analyzing the posts on social media, a person's emotions can be detected. This detection system can allow them to be aware of their mental health. This information can also play a momentous role in the decision-making process of a psychologist. For clinical depression, normally a psychologist evaluates his patients by taking a depression test based on questionnaires and academic interviews and recording it. But sometimes, it is not enough for detecting depression properly (Wang et al., 2013). Pointedly, these records are restricted because of many factors, such as sex, age, privacy, etc. To go beyond the boundary of clinical data, text mining tools extract and analyze depression syndrome from social media platforms such as Twitter, Facebook, and Instagram (Ma et al., 2017). Different techniques like ML classifiers, hybrid classifier models, and some new classifier models are proposed.

This paper is organized as follows: Section 2.1, states the background work regarding depression detection. A brief description of the proposed model is presented in Section 3.1. The result and performance are addressed by answering some research questions in Section 4.1. Section 5 concludes the paper by highlighting future works.

2. Background of the Study

Researchers are proposing different tools and techniques to detect depression from social media posts. Machine Learning techniques are frequently used in this area. The following section briefly discusses the related work in this field.

Mustafa et al. (2020) proposed an ML-based depression detection technique from Tweets. They used four ML classifiers such as Neural Network (NN), SVM, RF, and 1D Convolutional Neural Network (1DCNN) to build the model. They identified some keywords from the dataset and assigned weights to them. Subsequently, they matched the weighted words with previously generated fourteen psychological attributes in Linguistic Inquiry and Word count (LIWC) to classify those words into their respective classes of emotions. They classified them into three levels of depression (High, medium, and low) with an accuracy of 91%.

An AD prediction model for detecting anxious depression prediction in real-time tweets was proposed by Kumar et al. (2019). According to user posting features, they set five-tuple vectors such as words, timing, frequency, sentiments, and contrast. They developed their model using four ML classifiers such as Multinomial Naive Bayes (MNB), Gradient Boosting (GB), and RF, and Voting (majority voting) classifier which gave 85.09% of accuracy. For detecting the targeted words and transforming them into vectors using one-hot encoding and Word Embedding including the Word2Vec method, Ma et al. (2017) used an ML-based depression detection technique. Koltai et al. (2021) targeted specific hashtags which indicated depression. They used different techniques like NLP, NN, Latent Dirichlet Allocation (LDA), and LSA to get negative, and positive parts for a post. For classifying emotions into six categories like happiness, sadness, fear, anger, surprise, and disgust by analyzing the social media posts Gaind et al. (2019) used two different approaches: NLP including textual features (emoticons, degree words, negations, part of speech, and grammatical analysis), and Machine Learning classification algorithms. Finally, they achieved 91.7% accuracy using J48, and 85.4% accuracy using the SMO classifier.

By analyzing the social media texts Hassan et al. (2017) determined the binary and multi-class sentiment classification. They did feature extraction using POS Removal of stop words unigram, stemming, negation checker, and sentiment analyzer. They made comparisons among SVM, Naive Bayes (NB), and Maximum Entropy (ME) classifiers. The comparisons among classifiers were based on depression measurements where the accuracy of SVM was 91%, NB was 83%, and ME was 80%. Wang et al. (2013) proposed an ML-based depression detection technique for Chinese text. They used Waikato Environment for Knowledge Analysis (Weka) to develop their model. They considered BayesNet, Trees (J48),

and Rules (Decision Table) classification techniques to detect depression. The average ROC of the three classifiers was 85%. Moreover, it was 80% acceptable for the psychologist to detect depressed users in SNS. Balabantaray et al. (2012) were concerned about opinion mining and sentiment analysis by Natural Language Processing (NLP) and text mining that deals with automated discovery and classified emotions into six categories such as positive, negative, fear, joy, surprise, hate, and disgust. They used an emotion classifier based on multi-class SVM kernels which converted the seam words into numeric data. They reported that the accuracy was 72.34%.

A novel supervised algorithm namely Sequential S3 (SS3) for early depression detection proposed by Burdisso et al. (2019). The SS3 algorithm takes less time to classify than the other individual classifiers for example SVM, MNB, and Neural Network. The F-Score and precision are 0.61 and 0.63 respectively. Detecting depression in Reddit social media Forum Tadesse et al. (2019) proposed an ML-based model. For feature extraction, they applied a combination feature such as LIWC dictionary, LDA topic, and N-gram. LDA was chosen to reduce the input of the text data and to extract topics (features) from the text. MLP classifier achieved the highest accuracy, which is 91%. Islam et al. (2018) proposed an ML-based model for detecting depression on Facebook. They used four classifiers DT, SVM, KNN, and Ensemble classifiers. All classifiers' accuracy was between 60 to 80 percent. Chiong et al. (2021) used seven different ML classifiers namely Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), Multilayer Perceptron (MLP), Bagging Predictors (BP), Random Forest (RF), Adaptive Boosting, and Gradient Boosting (GB) for detecting depression from Twitter posts. Among them, Gradient Boosting (GB) classifier achieved the highest performance, with an accuracy of more than 98%. A hybrid algorithm which is dual classification with the fusion of SVM, and Naive Bayes (NB) algorithm used by Smys and Raj (2021). They have reported that hybrid classifiers (SVM, and NB) brought them higher accuracy than single classifiers (SVM, DT, RT, and NB).

Researchers are still facing many challenges regarding depression detection from text. However, recent trends in depression detection have adopted different techniques to enhance the performance of the ML classifiers. But they have not given enough focus on the contextual meaning of the text. In this work, we have mainly focused on the contextual meaning of the text whereas other dimension reduction techniques focus only to extract features and reduce dimensions. It can bring out the internal meaning of the text that helps to train the model more efficiently. Hence, in this work, a dimension reduction technique namely Latent Semantic Analysis (LSA) has been adopted to enhance the performance of the classifiers. This technique can extract the contextual meaning of the text which

helps the classifiers to achieve higher accuracy. In the meantime, a comparative study is given among four-dimension reduction techniques namely Latent Semantic Analysis (LSA), Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Fast Independent Component Analysis (Fast ICA). This study will help new practitioners to understand the importance of the dimension-reduction technique and help them to decide which dimension-reduction technique would be suitable for their model.

3. Depression Detection Model

In this paper, a depression detection technique using ML classifiers has been proposed. At first, data pre-processing techniques are applied to clean, transform and reduce the dimension of data which tends the model to work efficiently. Henceforth, models are evaluated using the six ML classifiers. Then, the outcome is reported based on the performance of the individual classifiers. Figure 1 shows the proposed model. The details are described in the following.

3.1 Machine Learning Classifiers

In this work, six ML classifiers namely SVM, DT, KNN, PA, RF, and Bagging have been used. The classifiers have different features to determine the optimal solution. A brief description of the classifiers is given below:

i. Support Vector Machine (SVM): It utilizes statistical learning theory to give optimized solutions. It fits the given dataset which returns a hyper-plane named ’best fit’. This hyper-plane segregates the dataset into classes. Utilizing the hyper-plane new classes are mapped into higher dimensional space and predicted what the class will be (Evgeniou & Pontil, 1999).

ii. Decision Tree (DT): This classifier consists of a root, internal node, branch, and leaf. To predict the class label, it starts working from root nodes, where the root node indicates the best attribute of the given dataset (Sharma & Kumar, 2016). The dataset is split into subsets. It compares the root attributes with the internal node attributes that represent a branch. It continues until it obtains the predicted class label at the leaf node.

iii. Random Forest (RF): It is an ensemble Decision Tree classifier. It adds randomness to the given dataset when building an individual decision tree and aggregates all of them. RF searches for the best feature while splitting the nodes among random subsets. All these combinations offer a more accurate and stable predicted class label (Biau & Scornet, 2016).

iv. K-Nearest Neighbor (KNN): It is a non-parametric algorithm. It is a method of finding the distance between the class and unknown class

v. of the dataset. It searches for the nearest neighbor between the classes and picks the class which gets the most votes. Afterward, this class is labeled as a predicted class (Novakovic et al., 2016).

vi. Passive Aggressive (PA): It is an online learning algorithm that works with margin base concept. It responds passive to correct classifications but is aggressive to the wrong classification. It penalized the model if it got an incorrect prediction. The model will make changes if the prediction is wrong (Crammer et al., 2006). It updates the classifier, adjusts it into the model, and labels it as the predicted class.

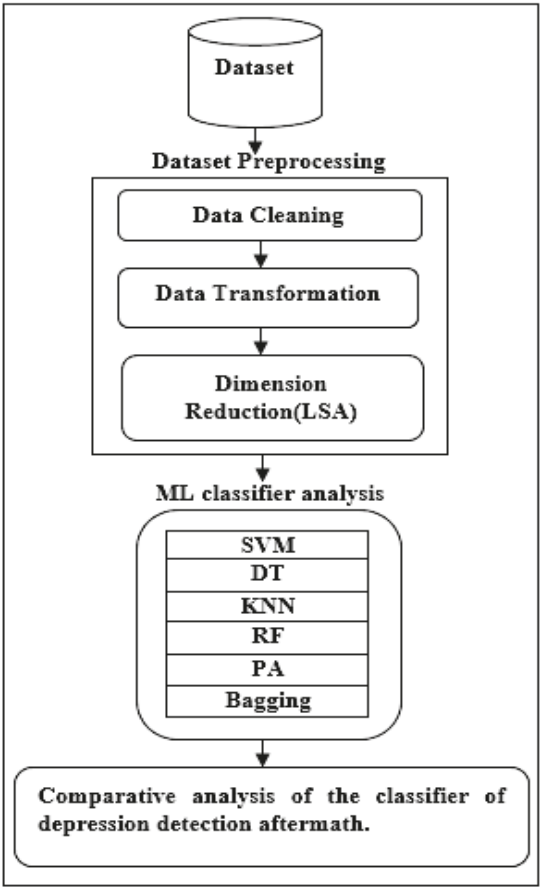


Figure 1: Flow diagram of depression detection technique using ML classifiers

vii. Bagging: It is an ensemble classifier that utilizes multiple models of the base classifier. It trains every model on a different set of data that

follows a technique: raw sampling with replacement (B uhlmann, 2012). Subsequently, it aggregates all the trained models using a voting method to build a stronger classifier whose predictive power is greater than the individual classifier. It reduces the variance of the dataset and gets rid of the trouble of over-fitting.

3.2 Data Pre-processing

In text analysis, data pre-processing is an important part as this can remove noise and unwanted elements from the text data and make the dataset more convenient to use. In this experiment, many data pre-processing techniques have been used for different purposes. They are discussed below.

3.2.1 Data Cleaning and Transformation

At first, URLs, duplicate words, extra spaces, and user mentions are removed to save time while analyzing the dataset. All the punctuations are erased to precise the data as it removes unnecessary signs which do not carry any information regarding depression. Then, all the stop-words are removed as they carry negligible information. For removing the suffix or prefix of a word or to find a root word, the Porter stemming technique has been applied. Further, tokenization is done with the RegexpTokenizer toolkit to split the tweets into words and then convert them into lowercase.

Frequency-Inverse Document Frequency (TF-IDF) vectorizer has been used to deal with the most frequent word in the dataset. It transforms the data with encoded numbers that carry the weights of the word and counts their frequent appearance. For contextual analysis, the uni-bigram model has been applied with the TF-IDF vectorizer.

3.2.2 Dimension Reduction Technique

For data pre-processing, dimension reduction is one of the most robust methods to reduce the data size as well as keep the variation of a dataset as much as possible. Among all the variations of data pre-processing, dimension reduction is the best choice for increasing the accuracy level because it avoids the over-fitting problem and takes care of multi-co-linearity. Besides, it helps with data visualization and is also useful for factor analysis.

i. Latent Semantic Analysis (LSA): In this work, A dimension reduction technique namely Latent Semantic Analysis (LSA) is used. It is also known as Latent Semantic Indexing (LSI). It reduces the dimension of the matrix. As LSA is language-independent, the dataset does not need to maintain the grammatical or auxiliary structure. While users are on social media expressing their thoughts, most of them are not maintaining sentence structure rules. LSA utilizes the sample vector directly, which

organizes the text structure semantically that helps the user's request for matching more accurately (Dumais et al., 1988) (Halko et al., 2011).

For example: instead of writing (I am a good kind of person), they write (I kinda good person). Although they are not properly making their sentences, we can easily extract the raw information from it through LSA. It does not depend on a specific word, string, or meaningful phrase. It can adapt to any kind of new, changing, or emerging thing. It is not sensitive to noise (unspelled data, arbitrary string) at all, any type of data can be read. It can perform context-based categorization (Landauer et al., 1998).

ii. Principal Component Analysis (PCA): Principal component analysis (PCA) is a method used to minimize the dimension of large datasets. It converts a huge quantity of variables into a small one and protects most of the information. It is like a characteristic expulsion method where we create new independent characteristics from the old. Initially, it normalizes the data. Thereafter, it calculates the covariance matrix. The next step calculates eigenvalues and eigenvectors. After completing the calculation, it chooses the components and forms the features vector. Finally, it forms principal components (Abdi & Williams, 2010).

iii. Latent Dirichlet Allocation (LDA): Initially, LDA concentrates on projecting the features in higher dimension space to the lower dimension. First, it is needed to calculate the separability between classes which means the distance between the mean of the different classes called between-class variance. Then, it calculates the distance between the mean and sample of each class called within-class variance. In the end, it fabricates the lower-dimensional space that can maximize the between-class variance and minimizes its within-class (Tran et al., 2019).

iv. Fast Independent Component Analysis (Fast ICA): It finds the latent topic in the text document. It presents all the hidden latent variables as linear combinations; those are statistically independent. In the beginning, it was used for signal processing but later it was found that it is also good for text analysis. The hidden latent variables are the text document topics, which provide the probability distributions on the universe of terms (Qi et al., 2001).

4. Result Analysis

To evaluate the effectiveness of detecting depression from social media (Twitter post), individual classifiers such as Decision Tree (DT), Support Vector Machine (SVM), K- Nearest Neighbour (KNN), Passive Aggressive (PA), Random Forest (RF), and Bagging have been used along with four-dimension reduction technique

namely Latent Semantic Analysis (LSA), Principal component analysis (PCA), Latent Dirichlet Allocation (LDA), and Fast Independent Component Analysis (Fast ICA). The result is reported based on four evaluation metrics namely precision, recall, accuracy, and F1-score. This work is conducted with an Intel Core i5 processor, 4GB RAM, 64-bit operating system, and Windows 10 education. Spyder (python 3.8) is used to implement models. Experimental details are discussed in the following.

4.1 Data Set for Depression Detection Model

The experiment is conducted with an existing Twitter dataset provided by Shen et al. (2017). They have collected their dataset from the Twitter API's. This dataset has been categorized into three parts namely D1, D2, and D3. D1 contains 292,564 depressed-related tweets of 1,402 users. D1 dataset has been collected from the tweets between 2019 to 2016. Here the tweets are marked as depressed when some words such as I'm, I was, I am found in the tweets. D2 contains 300 million users and 10 billion non-depressed related tweets, and it is fetched from the December 2016 posted tweets. The author labeled D2 dataset as non-depressed when the 'depress' word was never found in the user tweets. D3 consists of 36,993 depressed candidate users and more than 35 million tweets and these tweets are also from December 2016. Each dataset contains three subsets such as timeline, tweets, and users. The timeline subset consists of each user one monthly tweet post. The tweet subset contains unique tweets from the users. Finally, the user's subset contains information about each user. In this experiment, we have used the tweet subset. There are 6493, 5384, and 58900 tweets in D1, D2, and D3 tweet subsets respectively. The tweets are in different languages. D1, D2, and D3 tweets subset consist of 10, 78, and 91 types of language.

To train our model, 5000 tweets were randomly selected from both D1 and D2 datasets tweet subsets. Also, a similar number of depressed and non-depressed tweets has been collected from the D3 dataset tweet subset for testing our model. The D3 tweet dataset has been labeled with the **Table 1** lexicon words. **Table 1** presents 20 depression-related words. These words are the most frequent in the D1 dataset which is significantly related to depression. If the tweet contains any of these words, it is labeled as '1' means depression otherwise '0' means non-depression. Before comparing with **Table 1** there some pre-processing techniques are applied such as lower casing, hash-tag removal, and URL removal. After labeling it was found that there were 50,710 depression-related tweets and 8,190 normal tweet

Table-1: Word list for labeling of D3 dataset.

Abandon, abuse, problem, suffer, loser, fail, painful, depressed, depression, diagnosed, suicidal, broke, helpless, tired, torture, sick, ugly, insomnia, PTSD, destroy.

4.2 Performance Evaluation Metrics

Performance evaluation metrics mean the measurement of the quality of ML models using various measuring quantities. Evaluation metrics assure the legitimation and appeasement of a model (Botchkarev, 2019). For providing parallelism between the techniques, these metrics are used thoroughly. Here, four evaluation metrics are used namely precision, recall, F1-score, and accuracy. In this section, a descriptive study of these metrics is given.

i. True Positives (TP): It provides the number of states where the depression detection model can find out the real response to depression called true positive (T P) (Flach, 2003).

ii. True Negatives (TN): It presents the number of statues in which the model cannot find any depression, and no depression happens called the true negatives (TN) (Awoyemi et al., 2017).

iii. False Positives (FP): It indicates the number of times where the model detects depression, but it did not happen in the actual case (Zalpour et al., 2020).

iv. False Negatives (FN): It placed the falsely detected depression in the model, but it did not happen in the real case (Hemdan et al., 2020).

v. Precision (P): It provides the proportional positive measurement accurately (Pecorelli et al., 2019).

$$\mathrm{P} = \frac{TP}{TP + FP} \qquad (1)$$

vi. Recall (R): It grants the calculative result of actual positive value which are detected perfectly (Kurtanovi c & Maalej, 2017).

$$\mathrm{R} = \frac{TP}{TP + FN} \qquad (2)$$

vii. F1-Score (F): It approves the combination result of both precision and recall into a single meter (Ban et al., 2019).

$$\mathrm{F} = \frac{2 * P * R}{P + R} \qquad (3)$$

viii. Accuracy: The accuracy of an ML classifier means how it classifies data points accurately (Garc ıa et al., 2009). Accuracy refers to how many data point is perfectly predicted from all the data points. In other words, it is detected as the ratio of summation of the number of true positives and true negatives and the summation of the number of true positives, true negatives, false positives, and false negatives.

4.3 Research Questions and Evaluation

In this section, the model is empirically satisfied by addressing three questions RQ1, and RQ2, RQ3. RQ1 demonstrates how the classifiers perform to detect depression. RQ2 states how LSA influences the performance of the classifiers. RQ3 determines LSA is the best among these four-dimension reduction techniques. The following discussion covers the whole description of the above-mentioned research questions along with their assessment.

RQ1: How do the classifiers perform to detect depression from the text?

Among all six classifiers Bagging classifier performs the best in terms of accuracy whereas RF is the base classifier. This mixture helps to keep progressing the accuracy level. Bagging and RF both hold the bootstrap sampling feature and joining that similar feature makes the classifier more potential. The accuracy of the Bagging classifier with RF is 94.62%. **Table 2** shows that with Bagging, the RF classifier also achieves the highest precision level which is 0.95. Figure 2 shows the performance of the Bagging classifier with four different dimension reduction techniques.

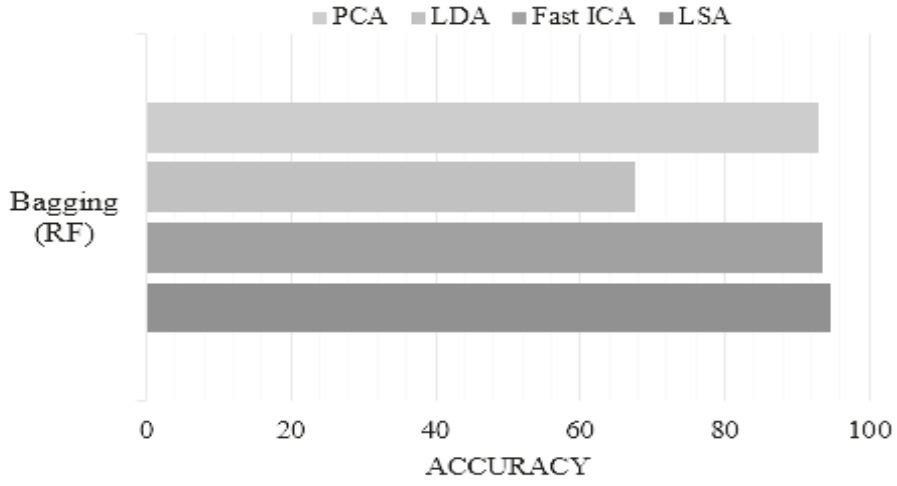


Figure 2: Performance of the dimension reduction techniques for Bagging classifier in terms of accuracy

Afterward, **Table 2** shows that ensemble classifier RF reveals a competitive result of 94.39%, as RF prevents overfitting problems with the help of multiple trees. **Table 2** shows that PA achieves the next highest result obtaining 93.88% accuracy. PA reacts passively for the right classification. KNN plays a substantial role as it is stable for the higher K number (k=8) for majority voting. Since KNN decreased the feature space taken as input, it obtains good accuracy.

Table-2: Performance of the classifiers in terms of accuracy, precision, recall, and F1-score with LSA.

Classifier	Accuracy (%)	Precision	Recall	F1-Score
SVM	92.11	0.92	0.92	0.92
DT	92.73	0.92	0.92	0.92
KNN	70.04	0.78	0.7	0.67
RF	94.39	0.94	0.94	0.94
PA	93.88	0.94	0.93	0.93
Bagging (RF)	94.62	0.95	0.94	0.94

Table 3 shows that KNN obtained 93.53% accuracy. The fifth best result is 92.73% which is achieved by the usual DT classifier since it minimizes the characteristics of a tree and gives a better prediction. SVM performs with an accuracy of 92.75%, while the kernel is Radial Basis Function (RBF). It executes the linear manipulations for mapping points into higher dimensional space, which makes it easier to separate the classification to make predictions.

Table-3: Performance of the classifiers in terms of accuracy, precision, recall, and F1-score with Fast ICA.

Classifier	Accuracy (%)	Precision	Recall	F1-Score
SVM	92.75	0.929	0.928	0.928
DT	92.09	0.921	0.921	0.921
KNN	93.53	0.937	0.935	0.935
RF	93.47	0.936	0.935	0.935
PA	89.51	0.909	0.895	0.894
Bagging (RF)	93.58	0.937	0.936	0.936

RQ2: How dimension reduction technique influences the performance of the classifiers?

Table 2, Table 3, Table 4, Table 5, and Table 6, illustrate the performance of LSA, Fast ICA, LDA, PCA, and without dimension reduction technique with ML

classifiers in terms of precision, recall, F1-score, and accuracy respectively. Dimension reduction is one of the most robust methods to reduce the data size as well as keep the variation of a dataset as much as possible. It takes less time to analyze the data.

Table-4: Performance of the classifiers in terms of accuracy, precision, recall, and F1-score with LDA.

Classifier	Accuracy (%)	Precision	Recall	F1-Score
SVM	62.48	0.643	0.625	0.613
DT	68.08	0.733	0.681	0.662
KNN	61.05	0.684	0.611	0.567
RF	67.38	0.759	0.674	0.645
PA	63.80	0.640	0.638	0.637
Bagging (RF)	67.50	0.763	0.675	0.640

By analysing all the dimension reduction techniques this paper came up with the decision that dimension reduction techniques have a great impact on the performance of the classifiers. The dimension reduction technique can boost the performance of the classifiers such as LSA and can decrease the performance, such as LDA.

Table-5: Performance of the classifiers in terms of accuracy, precision, recall, and F1-score with PCA**.**

Classifier	Accuracy (%)	Precision	Recall	F1-Score
SVM	90.81	0.921	0.908	0.907
DT	89.92	0.903	0.899	0.899
KNN	63.44	0.764	0.634	0.583
RF	93.13	0.934	0.931	0.931
PA	90.42	0.912	0.904	0.904
Bagging (RF)	92.90	0.932	0.929	0.929

Figure 3 shows that among all the classifiers Bagging performs best along with LSA which is 94.62%. **Table 6** shows that before applying reduction techniques, Bagging achieved an accuracy of 94.45%. **Table 3** illustrates that adopting Fast ICA also enhances the result in terms of accuracy as compared to without dimension reduction techniques though it performs less than LSA.

Fast ICA transforms the text into independent components, which makes it easier to separate the classification. Feeding PCA into the model does not give a good result. PCA does not work with finding the important patterns of the dataset all the

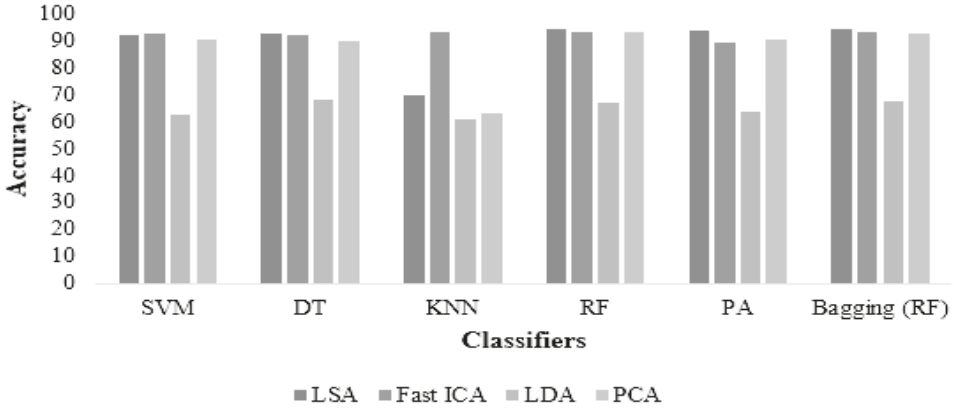


Figure 3: Performance of the dimension reduction techniques for ML classifiers in terms of accuracy

time. LSA always finds the important patterns of the dataset so that it gets the best prediction. LDA focuses on topic modeling whereas LSA emphasizes reducing the matrix size of a dataset while solving problems. **Table 4** and Figure 3 indicate that LDA as a reduction technique performs worse.

Table 2 shows that the accuracy of Bagging with RF is 94.62% whereas **Table 6** shows that, without the dimension reduction technique this drops down to 94.45%. **Table 6** shows that the accuracy of the PA classifier is 91.95%. **Table 2** shows that after adopting LSA, it becomes 93.88% and Table V shows that after adopting PCA it becomes 90.42%.

Table-6: Performance of the classifiers in terms of accuracy, precision, recall, and F1-score without any dimension reduction technique.

Classifier	Accuracy (%)	Precision	Recall	F1-Score
SVM	90.96	0.922	0.910	0.909
DT	94.01	0.942	0.942	0.940
KNN	58.51	0.756	0.585	0.502
RF	94.28	0.946	0.943	0.943
PA	91.95	0.926	0.919	0.919
Bagging (RF)	94.45	0.948	0.945	0.945

In SVM the best case of dimension reduction technique is Fast ICA. With Fast ICA, SVM also leads to greater accuracy which is 92.75%. Previously without any dimension reduction technique, it was 90.96%. According to the accuracy level, the performance of SVM, KNN, RF, and PA has increased after adopting LSA, and Fast ICA. The accuracy of KNN without any dimension reduction technique is 58.51%. The accuracy of KNN with LSA, Fast ICA, PCA, and LDA is 70.04%, 93.53%, 63.44%, and 61.05% respectively. LDA decreased all the classifier's performances except KNN. As KNN works to find the point in the nearest neighbor to make a prediction, Fast ICA also helps to find and separates the independent components of the dataset, so it greatly boosts the performance of KNN.

From the result analysis, the DT performs worst with the dimension reduction technique. Without the dimension reduction technique, it achieves an accuracy of 94.01% whereas with LSA, PCA and LDA, and Fast ICA. it achieves 92.73%, 89.92%, 68.08%, and 92.09%. **Table 6** shows that DT achieves the highest precision, recall, and F1-score level without any dimension reduction technique. As the dataset contains multi-languages it doesn't work great with LDA. LDA works worst when a single topic doesn't discuss coherently.

RQ3: Is LSA the best among these four-dimension reduction techniques?

Figure 3 demonstrates the comparison of classifiers with dimension reduction techniques in terms of accuracy. Among all the dimension-reduction techniques LSA does a tremendous job to attain accuracy with the maximum classifiers. Figure 4 shows the performance of classifiers in terms of accuracy where LSA is used as a dimension-reduction technique.

In this work, applying LSA makes a higher accuracy level possible with classifiers by reducing the dimension of the dataset. It understands the logic behind the text to classify tweets. It works fast as it uses less time and space to analyze users. **Figure 3** and **Table 2** shows that among all the classifiers Bagging performs best along with LSA which is 94.62%. With PCA, LDA and Fast ICA Bagging classifiers achieve accuracy of 92.9%, 67.5%, and 93.58% while the base classifier is RF. RF classifier achieved its highest accuracy of 94.39% with LSA. As RF and Bagging classifiers both have the bootstrapping technique, it helps LSA to interpret the text more correctly. Due to this, it achieves higher accuracy. PA also achieves its best accuracy, 93.88% with LSA. Though SVM, KNN, and DT don't achieve the highest accuracy with LSA, they gave a good performance with LSA. KNN, SVM, and DT achieve their second-highest accuracy of 70.04%, 92.11%, and 92.73% with LSA.

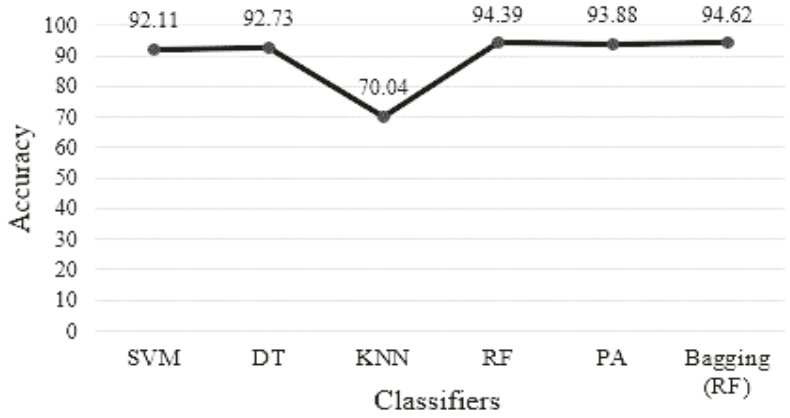


Figure 4: Performance of the classifiers in terms of accuracy with LSA

Comparing **Table 2, Table 3, Table 4, Table 5, and Table 6** shows that SVM and KNN do better with the Fast ICA algorithm. The Decision Tree does better without any dimension reduction technique. LSA also achieves the highest precision level of 0.95 with the Bagging classifier while the base classifier is RF.

Table-7: Performance of our and other authors' comparative proposed model on social media dataset.

Author Name	Classifiers	Accuracy (%)
Mustafa et al. (2020)	SVM, RF	91, 83
Kumar et al. (2019)	RF, Voting	81.04, 85.09
Gaind et al. (2019)	SVM, J48	91.7, 85.4
Hassan et al. (2017)	SVM	91
Proposed Model	SVM, RF	92.11, 94.39

The dataset contains multi-languages which makes it difficult for other dimension-reduction techniques to interpret the tweets. As LSA is language-independent, the dataset does not need to maintain the grammatical structure and it can interpret multi-language datasets. LSA removes the multi-collinearity of the classifiers. It also helps to represent the contextual meaning of the text. Due to all of these, it achieves higher accuracy with most of the classifiers. LSA helps to extract the raw information from the text which helps the algorithm to make correct predictions. After the study, it is proved that the dimension reduction technique (LSA) enhances the performance of the classifiers. Existing research works have used different techniques and ML classifiers. Most of the authors have used SVM and RF classifiers to develop their models. In **Table 7**, a comparison among some

existing research work has been given. It is noticeable that most of the researchers ignored dimension reduction techniques. In this work, SVM and RF with LSA perform 92.11% and 94.39% respectively in terms of accuracy. It is clear that LSA enhances the performance of the classifiers.

5. Conclusion

Most people over the world would not consult doctors at an early stage of depression because of negligence, and embarrassment. Besides, people share their feelings & emotions on social media platforms which is very helpful for detecting their mental health. In this paper, an ML-based depression detection technique from text is proposed. Four-dimension reduction techniques namely LSA, PCA, LDA, and Fast ICA are used to evaluate the performance of the algorithm. It is clear from the result that LSA and ICA help to increase the performance of the classifiers. With LSA, the Bagging classifiers perform the best in terms of accuracy while RF is used as the base classifier. The accuracy of the Bagging classifier with RF is 94.62%. LDA performs the worst among them. From the result analysis, it can be said that ML classifiers perform well to detect depression from the text. In the future, users' full profiles will be considered to detect depression. As we worked with only text-based analysis in this paper, in the future, user profiles, posted pictures, images, age, profession, and tweets will be analysed further to detect depression analysis more precisely. The deep learning model will be applied in the following version.

References

Abdi, H., & Williams, L. J. (2010). Principal component analysis. Wiley interdisciplinary reviews: computational statistics, 2(4), (pp. 433–459).

Awoyemi, J. O., Adetunmbi, A. O., & Oluwadare, S. A. (2017). Credit card fraud detection using machine learning techniques: A comparative analysis. In 2017 international conference on computing networking and informatics (iccni) (pp. 1–9).

Balabantaray, R. C., Mohammad, M., & Sharma, N. (2012). Multi-class twitter emotion classification: A new approach. International Journal of Applied Information Systems, 4(1), (pp. 48–53).

Ban, X., Liu, S., Chen, C., & Chua, C. (2019). A performance evaluation of deep-learnt features for software vulnerability detection. Concurrency and Computation: Practice and Experience, 31(19).

Biau, G., & Scornet, E. (2016). A random forest guided tour. Test, 25(2), 197–227.

Botchkarev, A. (2019). A new typology design of performance metrics to measure errors in machine learning regression algorithms. Interdisciplinary Journal of Information, Knowledge, and Management, 14, 45.

Buhlmann, P. (2012). Bagging, boosting and ensemble ¨ methods. In Handbook of computational statistics, Springer (pp. 985–1022).

Burdisso, S. G., Errecalde, M., & Montes-y Gomez, M. (2019). A text classification framework for simple and effective early depression detection over social media streams. Expert Systems with Applications, 133, (pp. 182– 197).

Chiong, R., Budhi, G. S., & Dhakal, S. (2021). Combining sentiment lexicons and content-based features for depression detection. IEEE Intelligent Systems, 36(6), (pp. 99–105).

Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S., & Singer, Y. (2006) Online passive aggressive algorithms.

Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., & Harshman, R. (1988). Using latent semantic analysis to improve access to textual information. In Proceedings of the sigchi conference on human factors in computing systems (pp. 281–285).

Evgeniou, T., & Pontil, M. (1999). Support vector machines: Theory and applications. In Advanced course on artificial intelligence (pp. 249–257).

Flach, P. A. (2003). The geometry of roc space: understanding machine learning metrics through roc isometrics. In Proceedings of the 20th international conference on machine learning (icml-03) (pp. 194–201).

Gaind, B., Syal, V., & Padgalwar, S. (2019). Emotion detection and analysis on social media. arXiv preprint arXiv:1901.08458.

Garc´ıa, S., Fernandez, A., Luengo, J., & Herrera, F. (2009). ´ A study of statistical techniques and performance measures for genetics-based machine learning: accuracy and interpretability. Soft Computing, 13(10).

Gupta, D., Bhatia, M., & Kumar, A. (2021). Real-time mental health analytics using iomt and social media datasets: Research and challenges. In Proceedings of the international conference on innovative computing & communication (icicc).

Halko, N., Martinsson, P.-G., & Tropp, J. A. (2011). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. SIAM review, 53(2), (pp. 217–288).

Hassan, A. U., Hussain, J., Hussain, M., Sadiq, M., & Lee, S. (2017). Sentiment analysis of social networking sites (sns) data using machine learning approach for the measurement of depression. In 2017 international conference on information and communication technology convergence (ictc) (pp. 138–140).

Hemdan, E. E.-D., Shouman, M. A., & Karar, M. E. (2020). Covidx-net: A framework of deep learning classifiers to diagnose covid-19 in x-ray images. arXiv preprint arXiv:2003.11055.

Hur, N. W., Kim, H. C., Waite, L., & Youm, Y. (2018). Is the relationship between depression and c reactive protein level moderated by social support in elderly? -korean social life, health, and aging project (kshap). Psychiatry investigation, 15(1), 24.

Islam, M., Kabir, M. A., Ahmed, A., Kamal, A. R. M., Wang, H., Ulhaq, A., et al. (2018). Depression detection from social network data using machine learning techniques. Health information science and systems, 6(1), (pp. 1–12).

Koltai, J., Kmetty, Z., & Bozsonyi, K. (2021). From durkheim to machine learning: Finding the relevant sociological content in depression and suicide-related social media discourses. In Pathways between social science and computational social science, Springer (pp. 237–258).

Kumar, A., Sharma, A., & Arora, A. (2019). Anxious depression prediction in real-time social data. arXiv preprint arXiv:1903.10222.

Kurtanovic, Z., & Maalej, W. (2017). Automatically classifying functional and non-functional requirements using supervised machine learning. In 2017 ieee 25th international requirements engineering conference (re) (pp. 490–495).

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. Discourse processes, 25(2-3), 259–284.

Ma, L., Wang, Z., & Zhang, Y. (2017). Extracting depression symptoms from social networks and web blogs via text mining. In International symposium on bioinformatics research and applications (pp. 325–330).

Murray, C. J., & Lopez, A. D. (1997). Alternative projections of mortality and disability by cause 1990–2020: Global burden of disease study. The lancet, 349(9064), 1498– 1504.

Mustafa, R. U., Ashraf, N., Ahmed, F. S., Ferzund, J., Shahzad, B., & Gelbukh, A. (2020). A multiclass depression detection in social media based on sentiment analysis. In 17th international conference on information technology–new generations (itng 2020) (pp. 659– 662).

Novakovic, J., Veljovic, A., Ilic, S., & Papic, M. (2016). Experimental study of using the k-nearest neighbour classifier with filter methods. In Proceedings of the conference on computer science and technology (pp. 91– 99).

Pecorelli, F., Palomba, F., Di Nucci, D., & De Lucia, A. (2019). Comparing heuristic and machine learning approaches for metric-based code smell detection. In 2019 ieee/acm 27th international conference on program comprehension (icpc) (pp. 93–104).

Qi, Y., Doermann, D., & DeMenthon, D. (2001). Hybrid independent component analysis and support vector machine learning scheme for face detection.

In 2001 ieee international conference on acoustics, speech, and signal processing. proceedings (cat. no. 01ch37221) (Vol. 3, pp. 1481–1484).

Safiri, S., Pourfathi, H., Eagan, A., Mansournia, M. A., Khodayari, M. T., Sullman, M. J., . . . others (2022). Global, regional, and national burden of migraine in 204 countries and territories, 1990 to 2019. Pain, 163(2).

Sharma, H., & Kumar, S. (2016). A survey on decision tree algorithms of classification in data mining. International Journal of Science and Research (IJSR), 5(4), 2094– 2097.

Shen, G., Jia, J., Nie, L., Feng, F., Zhang, C., Hu, T., . . . Zhu, W. (2017). Depression detection via harvesting social media: A multimodal dictionary learning solution. In Ijcai (pp. 3838–3844).

Smys, S., & Raj, J. S. (2021). Analysis of deep learning techniques for early detection of depression on social media network-a comparative study. Journal of trends in Computer Science and Smart technology (TCSST), 3(01), 24–39.

Tadesse, M. M., Lin, H., Xu, B., & Yang, L. (2019). Detection of depression-related posts in reddit social media forum. IEEE Access, 7, 44883–44893.

Tran, B. X., Latkin, C. A., Sharafeldin, N., Nguyen, K., Vu, G. T., Tam, W. W., . . . Ho, R. C. (2019). Characterizing artificial intelligence applications in cancer research: a latent dirichlet allocation analysis. JMIR Medical Informatics, 7(4).

Wang, X., Zhang, C., Ji, Y., Sun, L., Wu, L., & Bao, Z. (2013). A depression detection model based on sentiment analysis in micro-blog social network. In Pacific-asia conference on knowledge discovery and data mining (pp. 201–213).

Zalpour, M., Akbarizadeh, G., & Alaei-Sheini, N. (2020). A new approach for oil tank detection using deep learning features with control false alarm rate in highresolution satellite imagery. International Journal of Remote Sensing, 41(6), 2239–22.

Left Blank Intentionally

Fruits and Vegetables Disease Detection System Based on Indications Using Machine Learning Approach: A Systematic Review

Tajbia Karim[1], Mariam Chowdhury[2], Saima Murtuza[2], Afrida Israt Jahan[2], Afrina Khatun[3]

Abstract

In agriculture science, automated and computerized methods increase the country's growth, economy and productivity as it is highly dependent on the export of fruits and vegetables. Nowadays, it is impossible to check the quality of fruits and vegetables with bare hands as they are exported in a batch. In this world of technology, artificial intelligence plays an essential role by introducing many algorithms to detect diseases that hamper quality. This paper presents a detailed review of which algorithm best detects diseases in fruits and vegetables. The paper also includes details about pre-processing, segmentation, different algorithms for detection, and image enhancement. An analysis of different algorithms proposed by researchers for disease detection within fruits and vegetables was conducted. From contemporary research works, we have come to know that there is not one perfect method for detecting diseases of all fruits and vegetables. By careful analysis, we have recommended which machine learning method might be suitable for specific types of fruits and vegetables.

Keywords: Automated, Pre-processing, Segmentation, Fruits, Vegetables

1. Introduction

In this paper, we reviewed many papers that included diverse machine-learning approaches to detect diseases in fruits and vegetables depending on the indications. Bangladesh is an agricultural country, and this agriculture sector is crucial to the

[1] Department of Information and Communication Technology, Bangladesh University of Professionals, Dhaka, Bangladesh, Email: tajbia.karim@bup.edu.bd
[2] Department of Information and Communication Technology, Bangladesh University of Professionals, Dhaka, Bangladesh, Email: mariamchowdhury.200@gmail.com
[2] Department of Information and Communication Technology, Bangladesh University of Professionals, Dhaka, Bangladesh, Email: saima.murtuza13@gmail.com
[2] Department of Information and Communication Technology, Bangladesh University of Professionals, Dhaka, Bangladesh, Email: afridaisratjahan613@gmail.com
[3] Department of Information and Communication Technology, Bangladesh University of Professionals, Dhaka, Bangladesh, Email: afrina.khatun@bup.edu.bd

country's economy. The farmers of this country are highly dependent on producing fruits and vegetables. However, sometimes this production is hindered due to diseases caused by many pathogens. This is almost the same scenario for all fruits and vegetables. Researchers worldwide tried to give solutions to this specific problem described in their paper. By reviewing their works, we tried to analyze the best way to detect diseases among fruits and vegetables using different machine-learning approaches. As fruits and vegetables play a vital role in a country's economy as they are exported abroad, our analysis will be helpful for the researchers to find out which algorithm will be best for disease detection in fruits and vegetables with the highest accuracy rate, which will be beneficial for everyone, especially those who work in the farming sector and the industrial sector as well. The motivation behind this paper was to reduce the loss in the agricultural and food business due to fruit and vegetable diseases. As the weather in Bangladesh is changing every day, it is affecting agricultural products greatly. So, to reduce this loss, the researchers and other business personnel can follow this paper to eliminate fruit and vegetable diseases to a certain level.

By analyzing thoroughly, we found out that in many fruit businesses (export & import, juice companies, restaurants) these advanced techniques can be used to distinguish between fresh and rotten fruits and vegetables. In future research, if the researchers want to detect the maturity state of a fruit, it will become easy to find out what disease can infect it with the help of a practical machine learning approach. We compared datasets available in different papers and the approaches they followed. To the best of our knowledge, we could not find any papers that showed any comparison among the datasets and approaches. The comparison chart in our review paper reflects which algorithm is better for specific cases.

2. Literature Review

Researchers have started to apply Machine Learning to detect diseases in fruits and vegetables approximately since 2015 (Behera et al., 2018). The information about the work of researchers is shown in Table 1. The first model that we reviewed here was done in 2009 by (Qin et al., 2009). They proposed a model in which citrus canker was detected on citrus using hyperspectral reflectance imaging with spectral information divergence. Spectral Information Divergence (SID) was used for detection, with an accuracy rate of 96.2%. Behera et al. (2018) proposed a model using K-means clustering and Support Vector Machine (SVM) to detect diseases of mandarin and oranges. They noticed the symptoms of the disease in those plants by observing leaves, stems, and fruits. Also, they used image pre-

processing techniques such as image enhancement and lab color transformation. They were successful in detecting diseases with an accuracy of 90%. Then, Singh et al. (2019) proposed a system for detecting diseases in Mangifera indica (mango) by using Multilayer Convolutional Neural Network (MCNN). They used leaf for detection and for pre-processing, contrast enhancement and image reshaping were used. Their accuracy score was 97.13%. In the research of JJMCOE and J (2017), diseases were discovered in Malus domestica (apple). They have used the K-means clustering segmentation algorithm and Learning Vector Quantization Neural Network (LVQNN) for detection. For pre-processing, they have used image enhancement. They used 24 apple images that were real-time in nature, and they achieved an accuracy of 95%. Ullah et al. (2019) proposed a model to detect fruit disease by analyzing the leaves. They used contrast enhancement and image re-scaling for pre-processing and Convolutional Neural Network (CNN) as a detection algorithm. Their work had an accuracy rate of 97.33%. Ouyang et al. (2012) made a model to detect strawberry plant disease. The Top-hat transformation was used for pre-processing, which removes noise interference by median filtering. Their detection algorithm was an image segmentation algorithm with an accuracy rate of 99%. In the research of Pujari et al. (2013) grape, mango, and pomegranate fruits were used to grade and classify the disease of those fruits based on their statistical texture. Segmentation and Artificial Neural Networks (ANN) were used for detection purposes, and the accuracy rate was 84.65% for regular fruits and 76.6% for affected fruits. In 2019, (Doh et al.,2019) proposed a model using both SVM and K-Nearest Neighbour (KNN) for disease detection in citrus fruits and found an accuracy of 93.12% and 88.96% respectively.

In recent years, Yang et al. (2022) have achieved 99.70% accuracy in detecting disease in apples using a sorting image recognition system. Before that, Zaki et al. (2021) detected the disease in onion plants by observing young onion leaves. They have used CNN for this detection, and for pre-processing techniques, they have used image pre-processing. They achieved an accuracy of 77.05% for a data set of batch size 8 and 85.47% for a dataset of 16 images. Khattak et al. (2021) made a model which can detect infection in the citrus-based plant using their fruits and leaves and found an accuracy score of 94.55%. They have used CNN for this model. Another model where CNN is used as a detecting algorithm was proposed by Chowdhury et al. (2021). They did their research on tomato plants and found an accuracy score of 99.89%. Lastly, Mostafa et al. (2021) did their research on the guava plant using Deep Convolutional Neural Network (DCNN) structure and got an accuracy score of 97.74%.

3. Fruits and Vegetable Disease Detection

Disease detection of fruits and vegetables using Machine learning algorithms mainly involves five steps: Dataset collection, Data pre-processing, Segmentation, Feature selection, and Classifier.

3.1. Dataset

The primary data in fruit and vegetable disease detection applications are image data. Some of the main sources of collected data are google and various GitHub repositories. Some researchers also collected images from the fruits and vegetables available in the supermarket. Images acquired by different researchers are shown in Table 2.

3.2. Pre-processing

Images are acquired by various techniques consisting of multiple noises that worsen the aspect of an image. For this reason, the image without any pre-processing cannot contribute the appropriate data. Pre-processing enhances the image data, which overcomes reluctant misshaping and enlarges the image features essential for the processing and building an appropriate image (which is degraded form) than the original image for a specific application. The techniques used for image pre-processing for machine learning algorithms are Contrast enhancement and Rescaling.

3.2.1. Contrast Enhancement

The following equation is used to filter the contrast of any given image by assigning a constant intensity value to the pixel of that image using the histogram of an image. It is done with the help of the histogram equalization method.

$$H(P_{(x,y)}) = round\left(\frac{f_{(cdf)}(P_{(x,y)}) - f_{(cdf)min}}{(R \times C) - f_{(cdf)}} \times L - 1\right) \quad (1)$$

Here, $f_{(cdf)}$ = cumulative frequency of the gray level

$f_{(cdf)min}$ = minimum value of cumulative distribution function

$f_{(cdf)}(P_{(x,y)})$ = intensity of the current pixel,

R and C = product of several pixels in rows and columns

L = number of intensities

3.2.2. Rescaling

The images have been collected from different sources and captured with different devices. So, the images are not of the same size, therefore the images need to be resized.

3.3. Segmentation

Image segmentation is required to separate a digital image into distinct areas (Bhargava and Bansal 2021).

3.3.1. K-Means Clustering

K-means clustering is a type of vector quantization that divides n numbers of data into k number of clusters. Every observation is assigned to the cluster with the closest mean which acts as the prototype for the cluster. As a result, Voronoi cells are created in the data space (Vertica, 2022). Within-cluster variances are minimized using K-means clustering (squared Euclidean distances). Regular Euclidean distances, on the other hand, are not the more difficult Weber problem: the mean minimizes squared errors, but only the geometric median reduces Euclidean distances. Better Euclidean solutions can be produced, for example, by employing k-medians and k-medoids. The supervised KNN classifier, which is commonly confused with k-means due to its name, is connected to the unsupervised k-means algorithm. Using the cluster generated by k-means, the 1-nearest neighbor classifier is used to categorize incoming data into existing clusters.

3.3.2. Thresholding

Thresholding is an image segmentation technique that applies to the change of the pixels of an image to make it easier to analyze. It turns into a binary image from a color or grayscale image using thresholding. While the user can or should pick the threshold T manually in some circumstances, the user often prefers the threshold to be established automatically by an algorithm. The threshold should be the "best" threshold in certain circumstances, separating the brighter foreground items from the darker background objects. The threshold for picture intensity (image brightness) can be preset manually or automatically by some applications. Pixels whose bit values are zero turn black, whereas those with a bit value larger than zero become white (a bit value of one).

3.4. Feature Selection

Feature selection is the process of selecting the most important features to input in machine learning algorithms. Feature selection techniques are implemented to reduce the number of input variables by eliminating redundant or irrelevant features and narrowing down the set of features to those most relevant to the machine learning model. Some of the most common feature selection methods are:

3.4.1. Global Color Histogram (GCH)

GCH is a common term in machine learning. Color is regarded as one of the most important forms of visual representation. Changes to an image's size, rotation, or translation into different forms do not affect the image's color. Color space, color reduction, and the color feature extraction procedure are all factors to consider when extracting color characteristics from a picture (Han et al., 2022).

3.4.2. Color Coherence Vector (CCV)

CCV works by ranging each pixel as either coherent or incoherent. For each color, the number of coherent versus incoherent pixels is stored in a CCV. By separating pixels from incoherent pixels, it is possible to make finer distinctions than color histograms. They're all partially connected hidden layers, with the output layer being the fully connected layer at the end. The size of the input image is mirrored in the output shape (Pass et al., 1997).

3.4.3. SVM

SVM is a collection of supervised learning methods for classification, regression, and detecting outliers. The goal of the SVM algorithm is to find the best line or decision boundary for categorizing n-dimensional space into classes in the future so that new data points can be easily placed in the appropriate category. The best decision boundary is known as a hyperplane. Making a straight line between two classes is how a simple SVM works (McGregor and M, 2020). There are two types of SVM. Linear SVM is used for linear regression and classification problems. Kernel SVM is used for non-linear data. It can fit a hyperplane instead of a two-dimensional space with more features which has more flexibility for non-linear data.

3.4.4. Local Binary Pattern (LBP)

LBP is a robust method for image feature extraction in image processing. LBP reflects the correlation among pixels within a local area (usually a 3×3 area) which represents the local information (Bingham et al., 2015). By applying LBP, texture pattern probability can be precise in a histogram. LBP values need to be determined for all the image pixels (Prakasa, 2016).

3.4.5. Gray level co-occurrence matrix (GLCM)

It is a typical texture-based feature extraction model. The GLCM determines the relationship between pixels by performing an operation according to the second-order statistics in the images (Öztürk and Akdemir, 2018).

3.4.6. LVQNN

LVQNN is a superior classification technique for digital pictures. In general, it's a good idea to prepare data for LVQNN the same way it is done for KNN.

Table-1: Comparison of various classification methods for disease detection of vegetables and fruits

Work	Fruit/Vegetable	Affected Part	Pre-processing	Algorithm	Accuracy
Behera et al. (2018)	Mandarin	Leaf	Image enhancement	K-means clustering	90%
	Orange	Stem Fruit	Lab color transformation	SVM	
Zaki et al. (2021)	Onion	Young leaf	Image pre-processing	CNN	77.05% and 85.47% for a batch size of 8 and 16 dataset
Singh et al. (2019)	Mango	Leaf	Contrast enhancement Rescaling of the image data	MCNN	97.13%

JJMCOE and J (2017)	Apple	Fruit	Image enhancement	K-Means clustering segmentation algorithm LVQNN	More than 95%
Ullah et al. (2019)	Apple Bell paper Blueberry Cherry(sour) Corn(maize) Grape Orange Peach Potato Raspberry Soybean Strawberry Squash Tomato	Leaf	Contrast enhancement Image resizing	CNN	97.33%
Ouyang et al. (2012)	Strawberry	Fruit	Top hat transform Median filtering for removing noise interference	Image segmentation algorithm	99%
Pujari et al. (2013)	Grape Mango Pomegranate	Fruit	None	ANN classifier	84.65% and 76.6% for healthy and diseased fruits
Qin et al. (2009)	Citrus	Fruit	None	SID method for classification	96.2%

Khattak et al. (2021)	Citrus	Fruit and leaves	Keras image pre-processing using data generator class and API	CNN	94.55%
Yang et al. (2022)	Apple	Fruit	Grayscale processing Binarization Enhancement Processing Feature extraction	Sorting image recognition system	99.70%
Doh et al. (2019)	Citrus	Fruit	Image segmentation	ANN	SVM (93.12%)
			Feature extraction	SVM Phenotyping K-means clustering	ANN (88.96%)
Chowdhury et al. (2021)	Tomato	Leaves	Resizing and Normalizing Leaf segmentation Augmentation	CNN	99.89%
Mostafa et al. (2021)	Guava	Plant	Image acquisition Illumination	DCNN structures (AlexNet, SqueezeNet, GoogLeNet, ResNet-50, and ResNet-101)	97.74%

Table-2: Features of the datasets obtained by different researchers

Work	Fruit/Vegetable	Total Number of Image Data	Image Source
Ullah et al. (2019)	Apple	2,17,204	Internet
	Bell paper		Smartphone
	Blueberry		
	Cherry(sour)		
	Corn(maize)		
	Grape		
	Orange		
	Peach		
	Potato		
	Raspberry		
	Soybean		
	Strawberry		
	Squash		
	Tomato		
Singh et al. (2019)	Mango	2200	Real-time image
			GitHub repository
Zaki et al. (2021)	Onion	1000	Real-time image
JJMCOE and J (2017)	Apple	24	Real-time image
Khattak et al. (2021)	Citrus	2293	Different dataset
Chowdhury et al. (2021)	Tomato leaves	18,161	PlantVillage tomato leaf images

4. Comparative Analysis

After reviewing the works of multiple researchers, we can summarize that different algorithms work best on different fruits and vegetables. In Figure 1, we can get a brief idea about the accuracy rate for detecting disease in mango plants by using different algorithms.

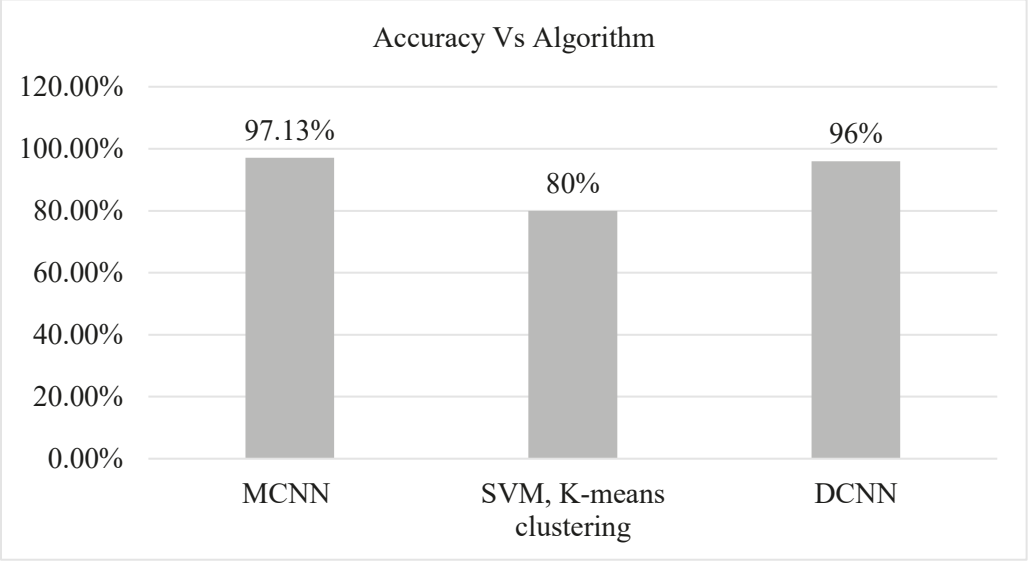


Figure 1: Accuracy vs algorithms for Mango disease detection

We have collected the data from the work of multiple researchers. Singh et al. (2019) used MCNN and found an accuracy score of 97.13%. Mia et al. (2020) proposed a model that used SVM and K-means clustering as detection algorithms and the accuracy rate was 80%. Ashok et al. (2021) used DCNN and found an accuracy of 96%. So, we can understand that MCNN has the highest score for accuracy in detecting disease in mango plants. By analyzing the comparison, we observed that MCNN shows exceptionally high accuracy in image recognition problems. It can also detect necessary features without any human guidance. On the other hand, for SVM and K-means clustering with an 80% accuracy rate, we observed if the dataset is large then SVM is not suitable as well as for K-means clustering, it is generally prone to biases.

We can conclude that MCNN is the best among the algorithms we analyzed due to its high accuracy rate and other features.

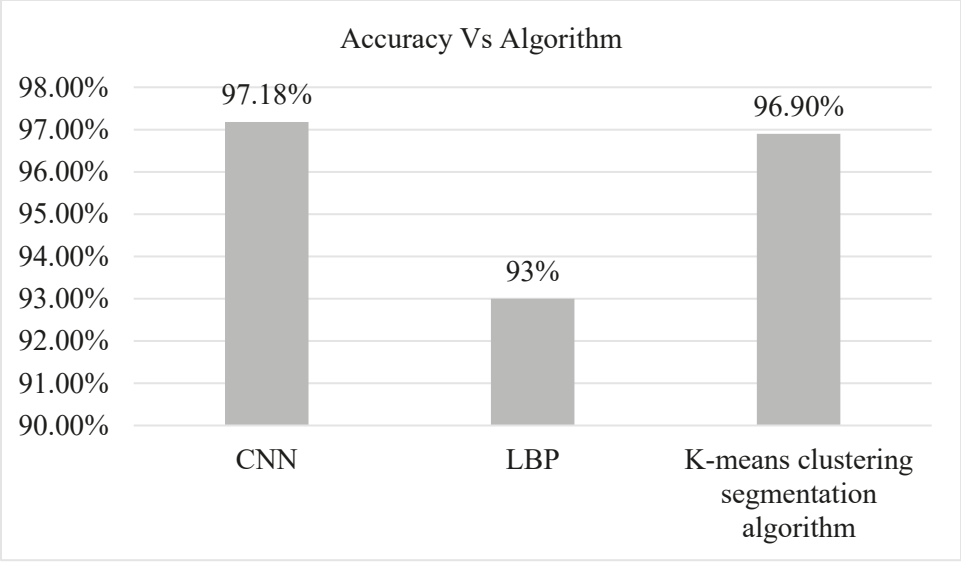


Figure 2: Accuracy vs algorithms for Apple disease detection

Figure 2 shows a bar chart of accuracy Vs different algorithms for detecting diseases in Apple. Khan et al. (2021) used CNN and found an accuracy score of 97.18%. 18%. Dubey et al. (2012) used LBP to complete the detection and found an accuracy of 93%. In the research of JJMCOE and J (2017), they used the K-means clustering segmentation algorithm and LVQNN and found an accuracy of 96.6%. So, we can say that the CNN algorithm has the best accuracy rate for detecting apple diseases. By analyzing the comparison of figure 2, we observed an accuracy rate of 97.18% for CNN. From the earlier comparison, we know that it gives a high accuracy rate in image recognition which is one of the notable features of CNN. In the figure, the LBP shows an accuracy rate of 93% which is the lowest among the other two algorithms. One of the disadvantages is its level of recognition is still low.

We can conclude that CNN is the best among the algorithms we analyzed due to its high accuracy rate and other features.

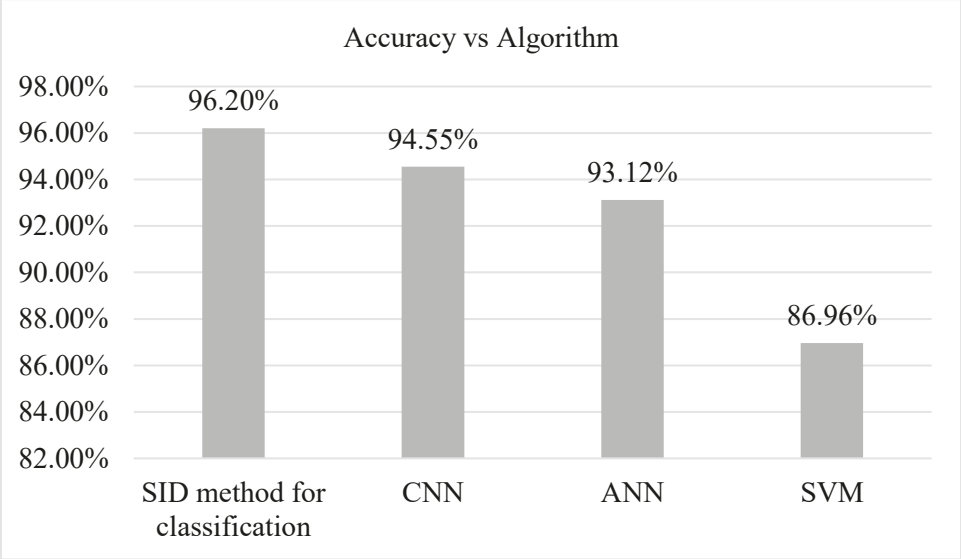


Figure 3: Accuracy vs algorithms for Citrus disease detection

Figure 3 represents a bar chart of different machine learning algorithms along with their accuracy rate to detect diseases in citrus fruit and leaves. From the above chart, we distinguished that the SID method has the highest accuracy rate. One of the major reasons for this value is that the SID method is an image-based classification procedure that is used to discriminate citrus diseases from other confounding diseases. On the other hand, SVM has the lowest accuracy rate among all other methods because it is sensitive to noisy attributes and requires a large search time.

We can conclude that for citrus disease detection SID method for classification is most suitable.

5. Conclusion

This paper highlights detailed descriptions of different algorithms for disease detection within fruits and vegetables along with applied pre-processing factors. The most significant features of agricultural by-products are their size, texture, and color. After reviewing and studying many research papers we came to a position where we can say which algorithm is best and will give high accuracy in detecting diseases in fruits and vegetables. More precise results can be obtained if researchers review papers with a large number of datasets which will allow them

to acquire higher accuracy rates and will help them to understand which machine learning algorithm is appropriate for detecting diseases in any kind of agro-based products. We can conclude that this comparative study of different machine learning algorithms can be applied to early-stage detection of commonly available fruits and vegetables preventing massive economic loss with an increase in GDP rate as a result of which any country can progress in doing business with agro-based products without any complication.

References

Ashok, V., & Vinod, D. S. (2021). A novel fusion of deep learning and android application for real-time mango fruits disease detection. In Intelligent System Design. Springer, Singapore, pp. 781-79.

Behera, S. K., Jena, L., Rath, A. K., & Sethy, P. K. (2018, April). Disease classification and grading of orange using machine learning and fuzzy logic. In 2018 International Conference on Communication and Signal Processing (ICCSP). IEEE, pp. 0678-0682.

Bhargava, A., & Bansal, A. (2021). Fruits and vegetables quality evaluation using computer vision: A review. Journal of King Saud University-Computer and Information Sciences, 33(3), pp. 243-257.

Bingham, E., Kaski, S., Laaksonen, J., & Lampinen, J. (Eds.). (2015). Advances in independent component analysis and learning machines. Academic Press.

Chowdhury, M. E., Rahman, T., Khandakar, A., Ayari, M. A., Khan, A. U., Khan, M. S., ... & Ali, S. H. M. (2021). Automatic and reliable leaf disease detection using deep learning techniques. AgriEngineering, *3*(2), pp. 294-312.

Doh, B., Zhang, D., Shen, Y., Hussain, F., Doh, R. F., & Ayepah, K. *(2019, September). Automatic citrus fruit disease detection by phenotyping using machine learning*. In 2019 25th International Conference on Automation and Computing (ICAC). IEEE, pp. 1-5.

Dubey, S. R., & Jalal, A. S. (2012, November). Detection and classification of apple fruit diseases using complete local binary patterns. In 2012 Third International Conference on Computer and Communication Technology. IEEE, pp. 346-351.

Han, Ju & Ma, Kai-Kuang. (2002). Fuzzy color histogram and its use in color image retrieval. IEEE transactions on image processing: a publication of the IEEE Signal Processing Society, 11, pp. 944-52.

JJMCOE, J. (2017). Detection and Classification of Apple Fruit Diseases using K-means clustering and Learning Vector Quantization Neural Network.

Khan, A. I., Quadri, S. M. K., & Banday, S. (2020). Deep learning for apple diseases: classification and identification. arXiv preprint arXiv:2007.02980.

Khattak, A., Asghar, M. U., Batool, U., Asghar, M. Z., Ullah, H., Al-Rakhami, M., & Gumaei, A. (2021). Automatic detection of citrus fruit and leaves diseases using deep neural network model. IEEE Access, *9*, pp. 112942-112954.

McGregor, M. (2020, July 02). SVM machine learning tutorial – what is the support vector machine algorithm, explained with code examples. Retrieved June 23, 2022, from https://www.freecodecamp.org/news/svm-machine-learning-tutorial-what-is-the-support-vector-machine-algorithm-explained-with-code-examples/ (Last accessed 21 July 2022)

Mia, M., Roy, S., Das, S. K., & Rahman, M. (2020). Mango leaf disease recognition using neural network and support vector machine. Iran Journal of Computer Science, 3(3), pp.185-193.

Mostafa, A. M., Kumar, S. A., Meraj, T., Rauf, H. T., Alnuaim, A. A., & Alkhayyal, M. A. (2021). Guava Disease Detection Using Deep Convolutional Neural Networks: A Case Study of Guava Plants. Applied Sciences, *12*(1), 239, pp. 1-19.

Ouyang, C., Li, D., Wang, J., Wang, S., & Han, Y. (2012, October). The research of strawberry disease identification based on image processing and pattern recognition. In International Conference on Computer and Computing Technologies in Agriculture. Springer, Berlin, Heidelberg. pp. 69-77.

Öztürk, Ş., & Akdemir, B. (2018). Application of feature extraction and classification methods for histopathological image using GLCM, LBP, LBGLCM, GLRLM and SFTA. Procedia computer science*, 132*, pp. 40-46.

Pass, G., Zabih, R., & Miller, J. (1997, February). Comparing images using color coherence vectors. In Proceedings of the fourth ACM international conference on Multimedia, pp. 65-73.

Prakasa, E. (2016). Texture feature extraction by using local binary pattern. *INKOM Journal, 9*(2), pp. 45-48.

Pujari, J. D., Yakkundimath, R., & Byadgi, A. S. (2013). Grading and classification of anthracnose fungal disease of fruits based on statistical texture features. International Journal of Advanced Science and Technology, 52(1), pp. 121-132.

Qin, J., Burks, T. F., Ritenour, M. A., & Bonn, W. G. (2009). Detection of citrus canker using hyperspectral reflectance imaging with spectral information divergence. Journal of food engineering, 93(2), pp.183-191.

Singh, U. P., Chouhan, S. S., Jain, S., & Jain, S. (2019). Multilayer convolution neural network for the classification of mango leaves infected by anthracnose disease. IEEE Access, 7, pp. 43721-43729.

Ullah, M. R., Dola, N. A., Sattar, A., & Hasnat, A. (2019, November). Plant Diseases Recognition Using Machine Learning. In 2019 8th International Conference System Modeling and Advancement in Research Trends (SMART). IEEE, pp. 67-73.

VerticaPy. (2022). Retrieved 22 June 2022, from https://www.vertica.com/python/documentation_last/learn/KMeans/ (Last accessed 21 July 2022)

Yang, M., Kumar, P., Bhola, J., & Shabaz, M. (2022). Development of image recognition software based on artificial intelligence algorithm for the efficient sorting of apple fruit. *International Journal of System* Assurance Engineering and Management*, 13*(1), pp. 322-330.

Zaki, M. A., Narejo, S., Ahsan, M., Zai, S., Anjum, M. R., & u Din, N. (2021). Image-based Onion Disease (Purple Blotch) Detection using Deep Convolutional Neural Network. Image, *12*(5), pp. 448-458.

Non-communicable Disease Detection Based on Early Symptoms Using Machine Learning Approach Enabling Smart Healthcare Model (IoMT)

Tasnim Binte Shiraj[1], Ali Mortuza[2], Kazi Md Anisur Rahman[3], Tajbia Karim[4]

Abstract

Early detection of disease can prevent fatality and even save the lives of individuals. Since many diseases may have some common symptoms, it is essential to critically analyze symptoms for the correct prediction of diseases. Machine learning influences disease prediction, analyzing numerous features with high accuracy. In our country, elderly people suffer mostly alone as every other member remains busy outside the home, so they lack proper care and constant observation. A cloud infrastructure that allows digital devices to gather, infer, and exchange health data is called the Internet of Medical Things (IoMT). As the global economy grows, so will the cost of linked healthcare. The ever-lowering cost of sensor-based technologies is the reason behind the extraordinary expansion of IoMT. This paper reviews which machine learning algorithm is most suitable for detecting non-communicable diseases in terms of precision, specificity, accuracy, and confusion matrix. It is possible to keep track of old persons by detecting disease from the early stages of symptoms. We used OHAS (Occupational Health Automated System) dataset for finding the accuracy of the disease detection system. We utilized several machine learning techniques for detecting non-communicable diseases (for example, K-Nearest Neighbor, Decision Tree, Support Vector Machine (SVM), XG-Boost, and logistic regression). This article's objective is to investigate the repercussions of using the aforementioned algorithms effectively and find out which is the best algorithm for early Disease detection. We observed that from the mentioned algorithms, XG-Boost outperforms all other algorithms and gives the best accuracy of 86.24 percent.

Keywords: Dataset, Classifier, Preprocessing, Feature Selection, XGBoost, SVM.

1. Introduction

A component of the Internet of Things is called the Internet of Medical Things (IoMT) which deals with the gathering, processing, data exchange, and storage for

[1,2,3,4]Department of Information and Communication Technology, Bangladesh University of Professionals, Dhaka, Bangladesh

Email: tasnimarrhiza512@gmail.com[1],alibinabdulwadud@gmail.com[2],anisurtopu555@gmail.com[3], tajbia.karim@bup.edu.bd[4]

medical purposes via a network of specific devices (things) designed to assure patient safety and data security. The IoMT platform is based on a common IoT platform. During the pandemic period, we face a major issue regarding medical treatment. It was really tough for us to visit the hospital and clinic on a regular basis. Also, various restrictions were a concern. To overcome this problem, we can build potable IoMT devices that can easily and accurately detect our diseases. On the other hand, many of us are not willing to go to the hospital for every single health issue. This type of negligence becomes the reason for critical health issues and diseases. In this circumstance, IoMT can play a big role. Like as we have a device that is able to detect diseases at a very early stage that might save one's life. It will help the patient to take proper decisions and clinical steps.

In recent times, various design options have been demonstrated using both existing protocols and regularly utilized bands. Various algorithms were utilized, including SVM, Logistic Regression, J48, KNN, Decision Tree, and Nave Bayes. Regression and classification are the two main applications of Support Vector Machine or SVM. Even when the data is not otherwise linearly separable, SVM categorizes data points by mapping them to a high-dimensional feature space. The data is processed such that the division may be shown as a hyperplane once the categories have been divided. Logistic regression is a technique for estimating the probability of a discrete output given an input variable. The most typical logistic regression model has a binary result.

In heart disease detection, using logistic regression, they get 87.1% accuracy. The supervised machine learning algorithm KNN (K-Nearest Neighbor) is used for classification and regression problems. The symbol "K" stands for the quantity of new unknown variables that must be predicted or classified and whose nearest neighbors must be counted.' As a result, the KNN algorithm can be employed in high-precision applications. The accuracy of the predictions is influenced by the distance measure. The KNN approach is best for applications that require a lot of domain expertise. J48 is a prominent machine learning method that can categorize and continually identify data. It consumes more memory and decreases the performance and accuracy of medical data classification when used, for example. J48 had the highest accuracy of 99.52 percent for identifying dementia in this literature review research. Although it may be used to solve classification and regression issues, the decision tree is most frequently employed to tackle classification difficulties. A probabilistic classifier with strong independence assumptions, the Naive Bayes method is based on probability models. This entire

technique is now very popular and is being used in a variety of studies. Doctors and medical institutions are also affected and are interested in these algorithms based IoMT devices for better treatment.

The main objectives of the paper are:

- To investigate the background and previous work in this field and find out which approach they took up and what they found in their research findings.
- To characterize the illnesses using various calculations such as XGBoost, K-Nearest Neighbor, Decision Tree, and Support Vector Machine.
- To find the most dangerous threat variables that cause these disorders.
- Comparison of various arrangement processes and identification of the most appropriate characterization strategy for provided data.

2. Literature Survey

Shokat Ali et al. (2020) presented Roles, Challenges, and Applications, which are based on the usage of IoMT guidelines and instruments has radically revolutionized orthopedic medical treatment, procedures, and services during the COVID-19 pandemic. Data gathering, report monitoring, patient database access, analysis of test pictures, and other tasks or features are all made possible by the IoMT system. Another issue to consider is interoperability. Rita Chhikara1 et al. (2018), proposed Analysis of Different Machine Learning Algorithms in Comparison for Dementia Detection. Dementia affects 46.8 million people now, with a projected 74.7 million more in 2030 and 131.5 million more in 2050. SVM, Logistic Regression, J48, and Nave Bayes were employed in this case. J48 has the highest accuracy of 99.52 percent. Good accuracy, specificity, discriminant power, and sensitivity are characteristics of the linear discriminant analysis said to Professor Pagar et al. (June 2021). They developed a model using the NB (Naive Bayesian), Decision tree map, Random Forest of Machine learning approach with the AES (Advanced Encryption Standard) algorithm, and SHDP (Smart Heart Disease Prediction) to address the issue of heart disease prediction. Heart disease may be identified with 87.1% accuracy using Logistic Regression. Diabetes prediction using Support Vector Machine (a linear kernel) was 85.71%, while Breast Cancer detection using AdaBoost classifier was 98.57 percent. Joyia et al. (2017) put forth The Applications, Benefits, and Future Challenges of Internet of Medical Things (IOMT) in the Healthcare Domain. The Internet of Things (IoT) is the most promising solution for the healthcare business, allowing individuals to control their own ailments and obtain aid in the event of an emergency. The Smart Rehabilitation System, Kidney Abnormality Detection System based on IoT Using Ultrasound Imaging, the Patient Physiological Monitoring System, Patient Posture

Recognition Application Using Supervised Learning, Safe and Intelligent Medical Healthcare System. Mohammad et al. (2020) suggested a healthcare monitoring system for the diagnosis of heart disease in the IoMT cloud environment using MSSO-ANFIS. This is based on the IoMT cloud environment's heart disease diagnostics. This employs using a neuro-fuzzy adaptive inference system (ANFIS) and MSSO (Modified Salp Swarm Optimization). MSSO-ANFIS enhances search capabilities by utilizing the Levy flight algorithm. ANFIS is a gradient-based learning algorithm that is prone to get stuck in local minima. MSSO-ANFIS has a greater accuracy of 99.45 percent with 96.54 precision than the other approaches. People may forecast heart disease by preprocessing (replacing missing values, data normalization) and feature selection. Taisa D et al. (2019) made a proposal called Breathing Monitoring and Pattern Recognition with the Worn Sensor, which is based on breath monitoring and pattern recognition using a wearable sensor. The nose, airways, lungs, etc. made the respiratory system. Many respiratory wearable sensors are used in this study, comprising acceleration sensors, oximetry sensors, humidity sensors, pressure sensors, and acoustic sensors. In this work, certain signal processing methods (amplification, filtering, analog to digital conversion, Fast Fourier transform (FFT)) are employed. This study employs Naive Bayes, SVM, and Artificial Neural Networks algorithms.

Danial H et al. (June 2017) proposed Diagnosis of Dementia from Early Clinical Data based on employing machine learning algorithms to identify dementia from clinical data. Based on the MMSE-KC data, data preprocessing will be performed. This study employs the Support Vector Machine (SVM), Naive Bayes, Logistic Regression, and bagging method. (SVM) is a data analysis and pattern recognition mapping model with an accuracy of roughly 87 percent. The FCM and PNN accuracy in this model were 74% and 69%, respectively. The Naive Bayes model was about 70% accurate. Bagging also aids in the reduction of variance and the prevention of overfitting. Using TriAxial Accelerometer Sensors, a Dementia Wandering Detection and Activity Recognition Algorithm was suggested by Kyu-Jin et al. (2009). This uses a triaxial method to identify dementia wandering where they offer an effective approach for translating raw sensor data into readable patterns in order to classify current activities, and then comparing these patterns to previously recorded patterns to detect many problematic patterns, such as wandering, one of the early indicators of dementia, and so on. The sensor node is made up of MSP430 for MCU, CC2420 for wireless connection, six 3-axis acceleration sensors, and three MUXs to select and receive data from sensors. Haruka et al. proposed An Early Detection System for Dementia Using the M2M/IoT Platform (2016). This system recognizes the signs of old people living

alone by using the M2M (Machine-to-Machine) / IoT (Internet of Things) platform. The authors devised three different sorts of analytical procedures. One way was behavioral data analysis/group comparison. Secondly, the way was behavioral data analysis/comparison with past behaviors. Another way was behavioral data and attribute data analysis utilizing Discriminant Analysis Method and Multiple Logistic Regression.

Radhika et al. (2020) discussed the use of decision trees and electronic health record analysis to predict diseases based on symptoms where they attempt to forecast ailments of users based on their symptoms. They employed the Decision Tree Classifier to attain their goal, which aids in detecting the patient's health state after obtaining their symptoms by providing the anticipated illness. To offer a summary of the health record, they employed NLTK - Natural Language Toolkit libraries. Their suggested system comprises two modules: one for illness prediction and the other for health records. Muhammad et al. (2021) observed the efficiency with which the smart healthcare (SHC) model can be used to monitor older persons testing aboard the IoMT dataset. Their proposed model used a way to monitor elderly persons using artificial neural networks (ANNs) which achieved 0.936 accuracy. Recently, there has been a surge of interest in incorporating machine learning into CPS, which can aid in disease categorization, detection, monitoring, and prediction for a variety of NCDs. As an example of NCDs, an original machine learning-based CPS for health is provided that successfully analyses data from wearable IoT sensors for diabetes risk prediction early on. Following tests with several machine learning methods, the researchers discovered that the Random Tree approach has the highest precision, needs the least amount of time to develop a model, and has a 94 percent accuracy in predicting the likelihood of diabetes at an early stage (Rahatara et al. (2021). A model for predicting multiple diseases based on Symptoms using Machine learning was proposed by the authors Talasila et al. (2021). The goal of this effort was to use Machine Learning (ML) models to help clinicians predict and analyze diseases at an early stage. They utilized a dataset that includes 4920 patient records that were judged to have 41 diseases. This research investigates three models' execution of a clinical record, yielding Decision Tree (0.973154), LightGBM (0.973154), and Random Forest (0.98315). Rinkal et al. authors created a technique for predicting diseases based on numerous machine learning techniques. There were more than 230 diseases in the dataset that was used for processing. They examined the dataset using Gaussian Nave Bayes, Fine, Medium, and Coarse KNN, Weighted KNN, Fine, Medium, and Coarse Decision Trees, Sub Space KNN, and RUSBoosted trees are among the ML models used. Among them, Weighted KNN gave the best result for the disease prediction model which had an accuracy of 93.5%. The authors Hamsagayathri et al. searched for the best ML techniques that are found to detect different specific

diseases. They observed various Computer Aided Diagnostic tests to examine ML techniques which results better. The authors found The Nave Bayes technique can be used to diagnose diabetes. SVM is useful for accurately diagnosing heart disease with an accuracy of 94.60%. For Liver Disease K Star resulted best with 83.47% accuracy and for Dengue disease DT, ANN, and RS resulted best with 99.96% correctness. The suggested deep learning system, Laavanya (2019) was trained where it uses 26,000 samples per dataset and obtains a 99.7% accuracy rate. As a result of this research, they built an accurate, rapid identification of stress levels As a result of this research, The "Stress-Lysis" IoMT system was developed, which may measure user-end stress levels (at the edge) and save the data in the cloud. To measure stress levels in real-time, the suggested Sensor device for stress and lysis may be readily fitted into a palm band or a glove. The proposed idea behind the IoMT-based stress detection system can provide monitoring of both acute and chronic stress. Deep Neural Networks (DNN) or Deep Learning Models are employed in huge dataset pattern recognition applications. The illness prediction system Sneha et al. (2020) constructed using a Decision Tree classifier, Random Forest classifier, and Naïve Bayes classifier among some other machine learning techniques is demonstrated in this study paper. This research gives a thorough comparison of three algorithms' performance on a medical record, each with an accuracy of up to 95%.

The suggestion with an improved linear model, recursion increased random forest (RFRF-ILM), Chunyan et al. (2020) to diagnose heart disease is described in this study. The purpose of this article is to elicit the main aspects of the prediction of cardiovascular disease using machine learning techniques. To cluster datasets, Decision Tree (DT) feature variables and criteria are utilized. The classifier is then used to estimate the performance of each data set. Because they have a lower error rate, the best performing models are chosen based on the findings. The suggested RFRF-ILM approach is used to combine the characteristics of the linear model with the random forest. The RFRF-ILM predicts cardiac disease with good accuracy. The suggested approach reduces diagnostic costs and time while increasing treatment accuracy. Dhiraj et al. (2019) provided a general illness prediction based on the patient's symptoms. They use the Machine learning techniques K-Nearest Neighbor (KNN) and Convolutional Neural Network (CNN) for accurate sickness prediction. The accuracy of CNN-based general sickness prediction is 84.5 percent, which is higher than the accuracy of the KNN approach. They begin by downloading a UCI's machine learning website which has a disease dataset. Following preprocessing, the feature was retrieved and chosen. The data is then classified using classification algorithms such as KNN and CNN. They can

accurately forecast illness using machine learning. A sample of 4920 patients' records with 41 disorders was chosen for study. 41 illnesses made up the dependent variable. 95 of 132 independent variables (symptoms) associated with illnesses were chosen and optimized. Laavanya et al. suggested a novel stress detection technique called iStress During Physical Exercise (2018), it measures stress levels by measuring body temperature, the pace of motion, and sweat. They constructed the Fuzzy type controller of the Mamdani type which gave accuracy as 97% accuracy in assessing a person's stress level. Peiying et al. (2019), proposed a system where the sentence structure is assessed by building a syntactic tree to extract the subject, predicate, and the object of the sentence from raw input sentences that have been evaluated by the syntactic analyzer to determine if the word sequence is lawful or not. Once the sentence has been preprocessed, it is sent into word2vec, which creates a vector representation of the sentence as input for CNN. Finally, they apply the Manhattan distance formula to calculate the sentence vector output's similarity score. They employed the SPO model to extract symptoms information as the neural network's input, then processed the model's output sentence vector using the Manhattan distance algorithm to generate the most similar disease prediction findings. The SPO model has a 75.8% accuracy rate, SPO outperforms previous models, demonstrating that it can better capture phrase meaning. Currently, medical applications employ body area networks, RFID, and Bluetooth. Blood pressure, ECG signals, and EMG activity are all monitored via sensors embedded in garments. With Optimized Neural Networks, Intelligent Guided Particle Local Search (IGPLONN) technique is suggested by Mohamed et al. (2020) based on IoMT. The suggested system's performance has been experimentally validated using MATLAB to ensure its superiority. When compared to existing approaches, the suggested IGPLONN based on the IoMT method achieves the highest accuracy of 98.3 percent. The system employs the backpropagation network throughout this procedure. The system's quality is measured using MATLAB results, which show that the suggested IGPLONN approach has the highest accuracy of 98.3% after comparing it to other approaches.

3. Framework of ML based Disease Detection

Figure 1: Data collection to result flow in ML framework

3.1 Data Collection

Data was gathered from the internet to diagnose the condition. Only the true symptoms of the condition were collected; no dummy values were used. The disease's symptoms were gathered from kaggle.com. The dataset that we used for this research article was OHAS Dataset. This csv file has 2129 rows of patient records with their symptoms (a variety of symptoms) and the condition that corresponds to them (148 a disease classification).

3.2 Data Pre-processing

Pre-processing data is a data mining technique that entails modifying or re-encrypting raw data to create an analysis-ready structure. It is known that efficiently the knowledge methodologies for pre-processing employed in provided work were deciphered using calculation as pre-processing of information, which included the following.

3.3 Data Purification

Certain measures are used to purify data. Along these ideas, settling the debt is similar to stuffing in lost value information with inconsistencies.

3.4 Data Scaling and Removing Redundancy

The dataset had some empty fields in some columns. We gave some random values to them (In the weight and height column) and some rows did not have symptoms which we removed for making the dataset workable. We also scaled the column's value.

The Kaggle repository provided the Disease/Symptoms database. It has numerous features (symptoms) and classes, as previously stated (diseases). To train the model, we use this to construct training and testing sets. We collect the user's symptoms and apply the trained algorithm to forecast the disease. A summary of the health report, on the other hand, is prepared based on the patient's record. Symptoms that are extremely important in relation to a specific condition. This is done in order to spread the sickness. In most cases, any prediction system will only look at a preset platform. However, with a low confidence level, this would only partially find the result. In order to arrive, we employ health record analysis in this paper to deliver individualized input, which offers us a higher level of confidence as well as system interaction with the user. We split the data of the dataset into two different sets. Training module set and testing the module set. We split this by a proportion of 80:20 which is a decent splitting done to design a prediction model. Disease prediction is one module of the total system. The other is linked to a medical record. The second lesson is intended to boost your confidence. To

achieve the best results, one of the training data sets was chosen. The user is involved in this process interface for gathering user input and informing them of their condition.

4. Methodology

The dataset we examined has a combination of stages that result in 148 diseases. This dataset contains 2129 documents from distinct patient samples, giving us a wide range of illness combinations of gender, height, weight, BMI Levels, Severity, and so on.

- The main goal of our approach is to enable elder citizens to easily access information about their health concerns without the assistance of technology or medical professional, even if they have no prior knowledge of the medical field.
- The fundamental idea behind how the algorithm generates more accurate projections is machine learning. The ANN or CNN or LSTM technique outperforms the medical field estimation in a comparable scenario.

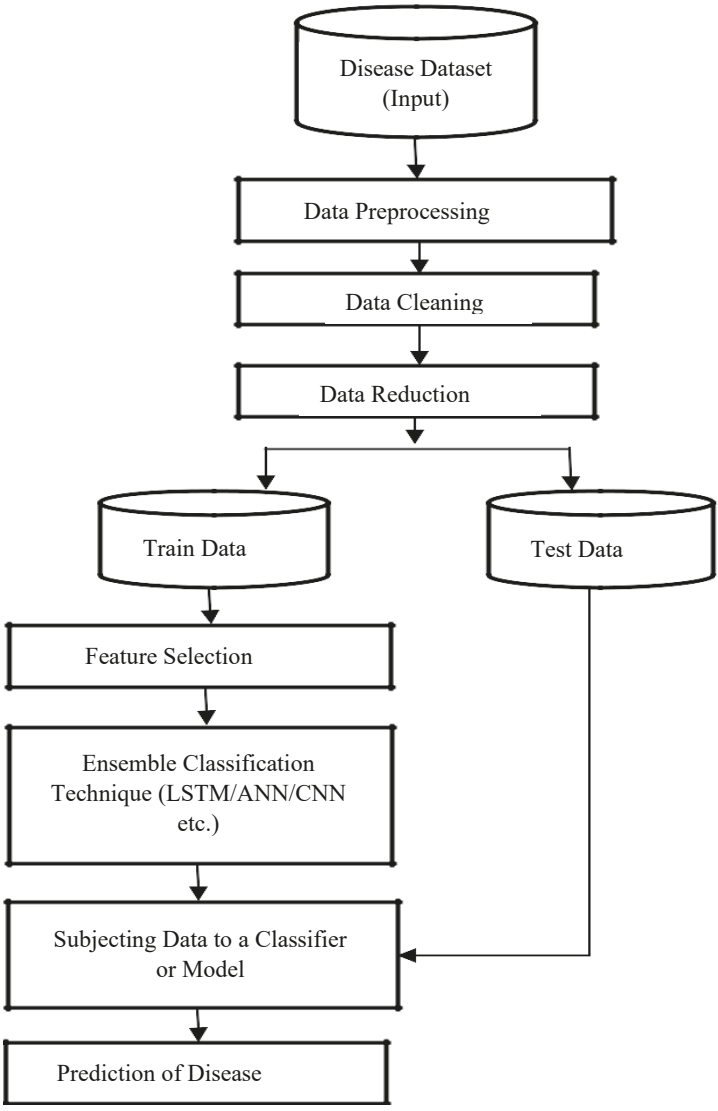


Figure 2: Methodology flowchart

A. Inputs (Patient Symptoms): When designing the algorithm, we assumed that the user would have a clear idea about the symptoms he is experiencing.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Disease	Disease_CUI	Symptoms	Symptom_CUI	Weight	Height	Intensity	Severity	Age	Gender	BMI_Leve	Region	Season
2	influenza	C0162565	uncoordi162ti	C0039239	68	180	high	medium	24	female	27.9	southwes	Summer
3	influenza	C0162565	fever	C0000737	68	170	low	medium	23	male	33.77	southeast	Summer
4	influenza	C0162565	pleuritic pain	C0235704	68	162	low	low	24	male	33	southeast	Summer
5	influenza	C0162565	snuffle	C0030554	68	162	high	medium	34	male	22.705	northwes	Summer
6	influenza	C0162565	throat sore	C0030552	68	185	low	high	21	male	28.88	northwes	Winter
7	influenza	C0162565	malaise	C0020538	68	185	medium	medium	21	female	25.74	southeast	Winter
8	influenza	C0162565	debilitation	C0020555	68	185	medium	medium	25	female	33.44	southeast	Winter
9	influenza	C0162565	asthenia	C0005758	68	185	low	high	25	female	27.74	northwes	Winter
10	influenza	C0162565	chill	C0030252	68	185	medium	low	27	male	29.83	northeast	Winter
11	influenza	C0162565	scleral icterus	C1868917	68	185	high	medium	27	female	25.84	northwes	Winter
12	influenza	C0162565	162sal flaring	C0392684	68	188	medium	high	31	male	26.22	northeast	Winter
13	influenza	C0162565	dysuria	C0012833	68	188	medium	low	31	female	26.29	southeast	Winter
14	influenza	C0162565	lip smacking	C0003467	68	188	high	high	31	male	34.4	southwes	Winter

Figure 3: Normal input data

B. Inputs (Preprocessed Dataset)

	A	B	C	D	E	F	G	H	I	J
1	Disease	Symptoms	Weight	Height	Intensity	Severity	Age	Gender	BMI_Leve	Season
2	influenza	uncoordi162tio	68	180	high	medium	24	female	27.9	Summer
3	influenza	fever	68	170	low	medium	23	male	33.77	Summer
4	influenza	pleuritic pain	68	162	low	low	24	male	33	Summer
5	influenza	snuffle	68	162	high	medium	34	male	22.705	Summer
6	influenza	throat sore	68	185	low	high	21	male	28.88	Winter
7	influenza	malaise	68	185	medium	medium	21	female	25.74	Winter
8	influenza	debilitation	68	185	medium	medium	25	female	33.44	Winter
9	influenza	asthenia	68	185	low	high	25	female	27.74	Winter
10	influenza	chill	68	185	medium	low	27	male	29.83	Winter
11	influenza	scleral icterus	68	185	high	medium	27	female	25.84	Winter
12	influenza	162sal flaring	68	188	medium	high	31	male	26.22	Winter
13	influenza	dysuria	68	188	medium	low	31	female	26.29	Winter
14	influenza	lip smacking	68	188	high	high	31	male	34.4	Winter

Figure 4: Preprocessed data

5. Implementation

Five classifiers are used to calculate information for the Disease Prediction Framework. These, for instance, were XG-Boost, Decision tree, K-Nearest Neighbor, Logistic Regression, and SVM. After that, we will try to create a model using ANN, CNN, or LSTM to classify diseases with a better accuracy rate than these classifiers.

5.1 Decision tree Model

The algorithm's order functioned as a Decision tree, which resembles a tree with multiple branches. As a result, it separates the dataset into smaller and smaller subgroups, resulting in the prediction of an objective value by examining the arrangement of unambiguous suppositions and then regulating on highlight esteem (manifestations for our circumstance) (disease). The most significant sections of a tree are the decision Node and the Leaf Node.

- Decision Node: A decision node is a node that has at least two branches. In our research, each manifestation is handled as a decision node.
- Leaf Node: The order is made up of leaf nodes, indicating that the decision can come from any branch. As a result, the node represents sickness on a tree.

5.2 XGBoost Model

An ensemble learning strategy is XGBoost. Relying just on the output of a single machine learning model might not always be sufficient. The prediction potential of several learners may be systematically combined through the use of ensemble learning. A single model that integrates the output of many other models is the ultimate result.

5.3 K-Nearest Neighbor

The k-nearest neighbors (KNN) technique is a data categorization method that uses the data points closest to it to estimate the likelihood that a data point belongs to one of two categories. To address classification and regression difficulties, the supervised machine learning technique k-nearest neighbor is applied. However, it is mostly employed to solve categorization issues.

5.4 Logistic Regression

The risk of categorization issues with two possible outcomes is modeled using logistic regression. It's an expansion of the classification issue with the linear regression model. When the dependent variable (target) is categorical, logistic regression is utilized.

As an illustration:

- To identify spam in an email (1) or (0)
- Whether or not the tumor is cancerous (1) or (0)

5.5 Support Vector Machine (SVM)

The Support Vector Machine, or SVM, is a well-liked Supervised Learning technique for dealing with classification and regression problems. The goal of the

SVM algorithm is to identify the ideal decision boundary or line for classifying n-dimensional space into groups so that the following data points may be quickly assigned to the appropriate category. A hyperplane is a mathematical term for the best decision boundary.

The extreme vectors and locations that will help build the hyperplane are chosen using SVM. Because support vectors are utilized as extreme instances, the method is known as an SVM classifier.

6. Results

After reviewing 24 papers we found several algorithms that used IoMT for early disease detection. For a hand on experience and examination, we chose the OHAS dataset and implemented Logistics Regression, Support Vector Machine, K-Nearest Neighbor, Decision tree, and XGBoost. We implemented these algorithms to understand the in-depth working procedure of mentioned algorithms. Nonetheless, when compared to the other models, XGBoost performs much better. Each model's accuracy score is listed below:

Table-1: Algorithms Vs Accuracy Score

Classifier/ Algorithm	Accuracy Score
XG-Boost	86.24%
KNN	70.276%
Decision Tree	54.95%,
Logistic Regression	58.164%
SVM (Linear Kernel)	68.92%
SVM (Polynomial)	65.58%

After preprocessing and data cleaning we imported the necessary libraries and implemented our dataset in different classifiers. We had a dataset containing 2129 rows from distinct patient samples, providing us with a huge variety of disease combinations based on gender, height, weight, BMI level, severity, and other factors. We may deduce from these results that each of the five Models performs excellently on the dataset.

7. Discussion

The paper's major goal was to look into the history and prior work on this subject to see which strategy they chose and what they discovered in their study findings. To describe the diseases using XG-Boost, Computations of K-Nearest Neighbor, Decision Tree, and Support Vector Machine. To identify the most serious risk factors that contribute to these disorders. Multiple arrangement methods are compared, and the best characterization methodology for the data supplied is chosen.

Data was gathered from the internet in order to diagnose the disease. There were no dummy values included, only the genuine symptoms of the disease were recorded. The signs of the disease were found on kaggle.com. The OHAS Dataset was the source of the data for this study. This csv file has 2129 rows of patient data, each of which includes their symptoms (a wide range of symptoms) and the condition that correlates to them (148 a disease classification). For processing the data, we followed three steps mainly, they are- Data Cleaning, Data Scaling, and Redundancy Removal.

The Kaggle source provided the Disease/Symptoms database. It has a range of traits (symptoms) and categories, as previously stated (diseases). This is used to create the model's training and testing sets. The symptoms of the user are collected, and the trained model is utilized to predict the ailment. On the other hand, depending on the patient record, a summary of the health report is prepared. Symptoms that are critical concerning a certain disease. This is done for the disease to spread. The dataset we worked on consists of a mix of phases that result in 148 diseases. In consideration of the 2129 documents of distinct patient samples in this dataset, we have a vast diversity of disease combinations with gender, height, weight, BMI Levels, Severity, and other variables.

For information calculation, the disease prediction framework uses five classifiers. These included XGBoost, Decision Tree, K-Nearest Neighbor, Logistic Regression, and SVM, to reference a few. We loaded the appropriate libraries and used our dataset in several classifiers after preprocessing and cleaning the data. This csv file dataset contains Symptoms, Diseases, Intensity, Severity, Gender, Season, etc. in Categorical form means the columns are in string representation, which is the Object datatype in the Pandas Data frame. Because categorical features are string data, humans can easily interpret them. Machines, on the other hand, are unable to assess categorical data rapidly. As a result, categorical data must be translated into machine-readable numerical data. Machine learning models are unable to interpret categorical data. As a result, the conversion to numerical representation is required. There are numerous methods for converting category

data to numerical data. To convert categorical data, we'll use the Label Encoding method.

We found the result for each implemented individual classifier which are, in XGBoost accuracy level was 86.24%, for KNN it was 70.276%, in the Decision tree algorithm it was, 54.95%, in Logistic Regression algorithm it was 58.164% and for the SVM linear kernel, the training accuracy was 68.92%, whereas test data achieved 64.86% accuracy. In the SVM Polynomial kernel, the training accuracy was 65.58% whereas we got 62.29% for test data accuracy.

8. Future Research Directions

Now we are trying to create a model using ANN, CNN, or LSTM to classify disease with a better accuracy rate than XGBoost, KNN, and Logistic Regression classifiers.

8.1 ANN

The learning methods used by artificial neural networks (ANNs) have the potential to adapt or learn on their own when new data is received. They, therefore, provide a great modeling tool for non-linear statistical data. ANN epochs (iterations) are made up of Forward and Backward Propagation.

The input layer receives information and passes it on to the hidden levels. Each input neuron is given weights and bias through the link between these two layers, and the weighted total, which combines the two, is then sent through the activation function. Each input neuron is initially given a random number of weights. The result is calculated by the Activation Function after it chooses which node to fire for feature extraction. This is known as forward propagation, Weights are updated after comparing with the original output, and error is known to minimize the error rate which is known as Backward Propagation.

8.2 LSTM

Recurrent neural networks known as Long Short-Term Memory (LSTM) networks are capable of learning order dependency in problems involving sequence prediction. This tendency is crucial in complicated problem areas like machine translation, pattern recognition, and other fields. This can be used in our model to recognize the same symptoms as a pattern and detect the correct disease every time and diagnose patients with what can be done for that specific disease. As we are willing to make an IOMT platform that can detect any disease from early symptoms, LSTM will be a great help in creating the model.

8.3 CNN

A convolutional neural network (CNN/ConvNet) is the form of a deep neural network used in deep learning to analyze visual pictures. Nowadays most of the people think neural networks to be matrix multiplications, but this is not the case with ConvNet. It uses a process known as convolution. A third function that specifies how the shape of one is changed by the other is produced by performing the mathematical process of convolution on two functions.

9. Conclusion

In this study, we review several papers and try out four approaches on a dataset to see how accurate they are. To diagnose the disease, data were acquired from the internet. The data for this study came from the OHAS Dataset. There are 2129 rows of patient data in this csv file. We primarily used three procedures to represent the data: Data Cleaning, Data Scaling, and Redundancy Removal. We worked with a dataset that includes a variety of phases that culminate in 148 illnesses, in light of the 2129 documents in this collection, including various patient samples. The illness prediction system employs five classifiers to calculate data. There was XGBoost, Decision Tree, K-Nearest Neighbor, Logistic Regression, and SVM, to mention a few. After preprocessing and cleaning the data, we loaded the required libraries and applied our dataset in different classifiers. We discovered that the accuracy level of the implemented classifier was 86.24 percent for XGBoost, 70.276 percent for KNN, and 58.164 percent for Logistic Regression.

References

Singh, R. P., Javaid, M., Haleem, A., Vaishya, R., & Ali, S. (2020). Internet of Medical Things (IoMT) for orthopaedic in COVID-19 pandemic: Roles, challenges, and applications. Journal of Clinical Orthopaedicsand, Trauma 11(4), 713-717.

Bansal, D., Chhikara, R., Khanna, K., & Gupta, P. (2018). Comparative analysis of various machine learning algorithms for detecting dementia. Procedia computer science, 132, 1497-1502.

Yede, N., Koul, R., Harde, C., Gaurav, K., & Pagar, C. S. (2021). General disease prediction based on symptoms provided by patient. Open Access International Journal of Science & Engineering, 6.

Joyia, G. J., Liaqat, R. M., Farooq, A., & Rehman, S. (2017). Internet of medical things (IoMT): Applications, benefits and future challenges in healthcare

domain. J. Commun., 12(4), 240-247.

Khan, M. A., & Algarni, F. (2020). A healthcare monitoring system for the diagnosis of heart disease in the IoMT cloud environment using MSSO-ANFIS. IEEE Access, 8, 122259-122269.

So, A., Hooshyar, D., Park, K. W., & Lim, H. S. (2017). Early diagnosis of dementia from clinical data by machine learning techniques. Applied Sciences, 7(7), 651.

da Costa, T. D., Vara, M. D. F. F., Cristino, C. S., Zanella, T. Z., Neto, G. N. N., & Nohama, P. (2019). Breathing monitoring and pattern recognition with wearable sensors. In Wearable Devices-the Big Wave of Innovation. IntechOpen.

Kim, K. J., Hassan, M. M., Na, S. H., & Huh, E. N. (2009, December). Dementia wandering detection and activity recognition algorithm using tri-axial accelerometer sensors. In Proceedings of the 4th International Conference on Ubiquitous Information Technologies & Applications (pp. 1-5). IEEE.

Radhika, S., Shree, S. R., Divyadharsini, V. R., & Ranjitha, A. (2020). Symptoms based disease prediction using decision tree and electronic health record analysis. European Journal of Molecular & Clinical Medicine, 7(4), 2060-2066.

Ishii, H., Kimino, K., Aljehani, M., Ohe, N., & Inoue, M. (2016). An early detection system for dementia using the M2 M/IoT platform. Procedia Computer Science, 96, 1332-1340.

Ferdousi, R., Hossain, M. A., & El Saddik, A. (2021). Early-stage risk prediction of non-communicable disease using machine learning in health CPS. IEEE Access, 9, 96823-96837.

Bhanuteja, T., Kumar, K. V. N., Poornachand, K. S., Ashish, C., & Anudeep, P. Symptoms Based Multiple Disease Prediction Model using Machine Learning Approach. International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN, 2278-3075.

Dahiwade, D., Patle, G., & Meshram, E. (2019, March). Designing disease prediction model using machine learning approach. In 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC) (pp. 1211-1215). IEEE.

Rachakonda, L., Sundaravadivel, P., Mohanty, S. P., Kougianos, E., & Ganapathiraju, M. (2018, December). A smart sensor in the iomt for stress

level detection. In 2018 IEEE International Symposium on Smart Electronic Systems (iSES)(Formerly iNiS) (pp. 141-145). IEEE.

Khan, M. F., Ghazal, T. M., Said, R. A., Fatima, A., Abbas, S., Khan, M. A., ... & Khan, M. A. (2021). An iomt-enabled smart healthcare model to monitor elderly people using machine learning technique. Computational Intelligence and Neuroscience, 2021.

Keniya, R., Khakharia, A., Shah, V., Gada, V., Manjalkar, R., Thaker, T., ... & Mehendale, N. (2020). Disease prediction from various symptoms using machine learning. Available at SSRN 3661426.

Hamsagayathri, P., & Vigneshwaran, S. (2021, February). Symptoms Based Disease Prediction Using Machine Learning Techniques. In 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV) (pp. 747-752). IEEE.

Khan, M. A., & Algarni, F. (2020). A healthcare monitoring system for the diagnosis of heart disease in the IoMT cloud environment using MSSO-ANFIS. IEEE Access, 8, 122259-122269.

Hashem, M., Vellappally, S., Fouad, H., Luqman, M., & Youssef, A. E. (2020). Predicting neurological disorders linked to oral cavity manifestations using an IoMT-based optimized neural networks. IEEE Access, 8, 190722-190733.

Rachakonda, L., Mohanty, S. P., Kougianos, E., & Sundaravadivel, P. (2019). Stress-lysis: A DNN-integrated edge device for stress level detection in the IoMT. IEEE Transactions on Consumer Electronics, 65(4), 474-483.

Guo, C., Zhang, J., Liu, Y., Xie, Y., Han, Z., & Yu, J. (2020). Recursion enhanced random forest with an improved linear model (RERF-ILM) for heart disease detection on the internet of medical things platform. IEEE Access, 8, 59247-59256.

Grampurohit, S., & Sagarnal, C. (2020, June). Disease prediction using machine learning algorithms. In 2020 International Conference for Emerging Technology (INCET) (pp. 1-7). IEEE.

Zhang, P., Huang, X., & Li, M. (2019). Disease prediction and early intervention system based on symptom similarity analysis. IEEE Access, 7, 176484-176494.

OHAS:https://www.kaggle.com/code/kerneler/starter-disease-prediction-through-112c9798-a/data?select=OHAS+Dataset.csv

Present Situation Analysis of Environmental Management System on a Proposed LEED GOLD Sweater Manufacturing Company

Umma Salma Hashe[1]

Abstract

The textile industry has become the backbone of the economy of the country, but recently it also has risks. Large businesses on the higher end of the value chain are increasingly placing pressure on small and medium enterprises to satisfy minimum environmental and social requirements with quantity, quality, and price-cost requirements. Environmental management is the management of the responsibility of an organisation for its environmental impacts. The proposed LEED GOLD sweater manufacturing company is a brand-new subsidiary of the Rashid Group. The aim of this study is to examine the current situation of sweater manufacturing company's environmental management system and to find out the shortcomings of Environmental management system practices. In this analysis, two data collection strategies were adopted. One of them is field based study along with questioner survey and another one is collection assisted by official paper, literature review, etc. This study found that general environmental policies and approaches to management are strictly defined in this factory. The research demonstrates that strict monitoring and compliance of existing laws led the owners of the factory to prevent a poor system of environmental management. In addition to the problem of emissions, factory owners strictly comply with environmental laws and pollution regulations except for wastewater discharge management. Furthermore, the study results and recommendations will be of great benefit to the authority, including HR & Environmental Enforcement Management, EMS Management, which would show best practices and pathways for the industry to help achieve LEED GOLD certification in order to minimize its environmental impacts and improve its operating efficiency.

Keywords: Environmental management system, Environmental policies, Environmental laws, Environmental compliance checklist, Environmental Enforcement Management, Audit

---

[1]Department of Environmental Science, Bangladesh University of Professionals, Dhaka, Bangladesh

1 Introduction

The manufacturing industry can no longer neglect environmental concerns. Companies face increasingly strict regulatory restrictions and rising prices of utilities and raw materials. For company to thrive in a competitive market, the efficient and successful use of raw materials and enhanced process operations are essential. Suppliers and clients are now exerting pressure on businesses to reduce their negative environmental effects. All of these considerations mean that an important aspect of the business strategy should be environmental concerns. Setting up an environmental management system (EMS) would provide a company with a structure by which it can monitor and enhance its environmental efficiency. An EMS will enable to define opportunities by reducing waste generation and reducing the use of water and other utilities to minimize operational costs. According to Khanna and Anton (2002) EMS represents an organizational transformation within companies and a self-motivated attempt to rationalize environmental externalities through the implementation of management practices that combine environmental and production decisions that recognize pollution mitigation opportunities and enable the business to make continuous improvements in production methods and environmental performance. Clement (1996) notes that, the standard refers to those environmental factors over which the business has control, or which may be expected to affect the company. Aboulnaga (1998) pointed out that, the adoption and use of an EMS will provide companies and organizations that want to compete on the international stage a competitive advantage. Other proponents of ISO 14001 like Stapleton et al (2001) argued that the requirement could act as a framework to enhance organizational performance significantly. A company will increase its market share and reduce expenses related to environmental taxes, electricity, waste, water use, pollution, and fines by adopting and being accredited. Bangladesh embarked upon rapid industrialization after its independence from Pakistan in 1971. The establishment of industries for the country's rapid economic growth is being promoted. In urban areas such as Dhaka, Narayanganj, Narshingdi and Chittagong, most of these factories are located (BBS, 2010). A number of environmental concerns have now been identified and traced to industrial facilities in Bangladesh. The four major rivers near Dhaka Shitalakhya, Balu, Turag and Buriganga receive more than 1.5 million cubic meters of wastewater every day from the surrounding industrial units, according to a World Bank report. The report demonstrates that the improper treatment of wastewater by a number of factories results in the release of highly toxic waste causing

significant harm to the atmosphere and to marine life. More than 30 to 40 lakes suffer from the effects of poor water quality exacerbated by untreated industrial waste in the river system. The proposed LEED GOLD sweater manufacturing company is a brand-new Rashid Group subsidiary. They have existing environmental management system with have some deficiencies. They have an existing system for environmental management with certain deficiencies. The study will help to understand the present situation of environmental management system of proposed LEED GOLD sweater manufacturing company along with find out the drawbacks of EMS practices. It will be of great use to the authority, including HR & Environmental Enforcement Management, EMS Management, to take measures in future research and development and in all decision making on environmental management system concerns.

2 Methodology

2.1 Study Area

The study was conducted on a proposed LEED GOLD sweater manufacturing company, which has the most advanced sweater manufacturing system in the RMG sector of the country and manufactures the highest quality sweaters for the world's leading brands. The facility is a member of ACCORD and other compliant bodies concerned. The 110,000sft factory is fitted with 600 sets of automatic jacquard knitting machines and other branded machinery of the Shima Seiki, Gousheng & Hongqima brand for linking to final finished goods. And the factory also has a sufficient number of specialists, trained & qualified staff to empower these machines. Thus, the technology combined with experience results in the factory providing valued customers with approx. 400,000 pcs of finished sweaters every month. This Factory goal is to well survive heightening the image of the group along with maintaining environmental management system. This factory called green factory because:

i. Sustainable Sites: The factory is conveniently located close to all important services, where school and market are located. The site also offers secure bicycle parking facilities and facilitates the use of non-fossil fuel transportation within 500 meters of the mosque and bus stops.

ii. Water Efficiency: Fresh, drinkable water is a precious commodity that is only available in a limited supply. Therefore, prudent use and treatment of

are very important as we look to expand our built environment and shrink natural treatment areas by reducing indoor & outdoor water use.25% water saving fixtures & fittings are used all over the premises along with proper flow meters to maintain use sand consumptions.

iii. Energy & Atmosphere: The factory building has been designed in a way which saves energy and renewable power sources are also added to the project.25KW solar power is infused on grid with the main power source. The project is using NONCFC (Chlorofluorocarbon) Evaporative Cooling Systems complying the perspective measures of ASHRAE Standard 90.12010. Hence resulting to optimized energy performance and enhanced refrigerant management to avoid global warming.

iv. Indoor Environmental Quality: The quality of the air on the inside of a building is important to all the occupants, as it can contribute to illness and lack of productivity. The factory design allows an increased amount of fresh air by using building materials and products without harmful chemicals to improve the air quality. It has a wide access to the outside views for providing day light and natural lighting and ample of space for better thermal comfort.

v. Innovation & Design: The innovative & exceptional designing of all environmental aspects the factory has been done by a professional LEED accredited personnel with the special ability for green buildings.

vi. Materials & Resource: What materials are used, where they come from, how they are made, and how they are disposed of are instrumental in determining how green a project is. The factory uses more green materials, including renewable materials, recycled materials, and natural materials.

2.2 Research Design

This research design was focused on the common understanding of the EMS model's control/corrective action in Bangladesh. Considering the Standard Policy of Bangladesh, environmental aspects such as: air emissions; waste generation and management; river wastewater releases; land contamination; use of energy and natural resources; Local and Community issues (Noise, odour, aesthetic condition); Positive aspects of operation on the environment and Environmental Impacts including Local air pollution; Global air pollution;

Adverse impact on employee health; Degradation of local water quality; Depletion of natural resources; Degradation of flora and fauna and Noise pollution. As shown in figure the backbone of the study is the "Existing Situation Analysis" which is macro scale analysis on environmental management system. At the end of the study the most significant results revealed from the study were listed briefly and recommendations were developed.

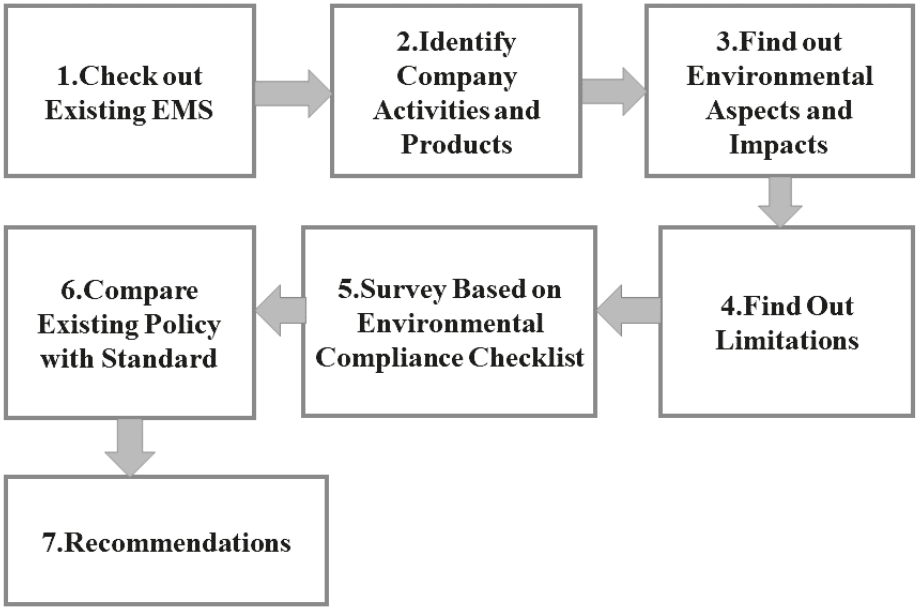


Figure 1: Research Design of this Study

2.3 Data Collection

A questionnaire survey and field survey were carried out to accomplish the study goals. At the outset of the research, a study design was developed to frame the study as well as to design the questionnaire and field survey. To collect the primary data from the questionnaire, a field survey was carried out. Then the data was analysed to obtain the results. Primary data was the main source of information for the study. Some secondary data sources have also been implemented. Books, official reports, published electronic and print journals, and information from local environmental experts were among the secondary data sources.

,

2.4 Questionnaire and Field Survey

The design of the questionnaire and field survey was based on the proposed LEED GOLD sweater manufacturing company environmental compliance checklist, including Environmental Management System; Waste Management; Wastewater and Effluent Management; Air Emissions Management; Water and Food Management; Energy Usage and Greenhouse Gas Management; Land Use and Biodiversity; Hazardous Substances and Noise Management. The standardized questionnaires were administered over the four hundred staff and management authority of proposed LEED GOLD sweater manufacturing company in order to carry out the survey most effectively. Some questionnaires were left with in which the questionnaire could not be administered personally. A convenient purposeful sampling technique was used for the field survey over the employees and authority to collect data from the proposed LEED GOLD sweater manufacturing company. Industry was chosen on the basis of their availability and accessibility to the author and their willingness to participate in the survey, including environmental enforcement management and employees from each production section.

2.5 Data Analysis and Presentation

The analysis was considered to be exploratory and did not statistically analyse the results. In terms of the general patterns that emerged from the study, the findings were then compared with the Standard Environmental Management System outlined. And, based on the results obtained from the study, recommendations have finally been developed.

3 Research Findings and Analysis

Research findings were found out from questionnaire and field survey that was administered over the four hundred staff and management authority of the proposed LEED GOLD sweater manufacturing company. The next sections illustrate the research findings.

3.1 Environmental Management System

The proposed LEED GOLD sweater manufacturing company is a green sweater factory and has an existing environmental management system to comply with environmental policy, minimize its environmental impacts and improve its operational performance. It has three storey buildings and different sections are apprehended by each building. Each section apprehends fire equipment, work-related procedure, emergency evacuation plan, section wise policy sheet, BSCI code sheet, leaving procedure, ETI base code, pest control, emergency plan, etc.

3.2 Temperature & Humidity Level

This factory has separated production area and all required Temperature and humidity level quality assessment carried out by green tech Environmental Company according to the CSA & OSHA standard. The Temperature and humidity level present within the acceptable limit in all areas.

3.3 Light Level Quality

The Green Tech Environmental Company checked the appropriate light quality parameter for separate production area in compliance with the Bangladesh Labor Laws, Bangladesh National Building Code & OSHA Standard. The light level quality found in all areas within the acceptable limit.

3.4 Policy

Seventy-five policies have been adopted by the proposed LEED GOLD sweater manufacturing company and among them forty policies relate to the environmental management framework, including environmental policy, health & safety policy, healthy chemical handling policy, fire safety policy, environmental risk assessment and management policy, etc., according to National Environmental Policy 1992, Environmental conservation Act-1995(Amended in 2000, 2002 & 2010); Environmental Conservation Rules-1997(Amended in 2002, 2003 & 2010); Industrial policy-2010; Labor act-2006 etc.

3.5 Audit

They carried out three types of audit schemes, including internal audit, external audit (buyer requirement basis) and requirement-based environmental audit.

3.6 Legal Document

They apprehend Certificate of incorporation, Factory License, Trade License, BGMEA Membership License, Group Insurance, Noc, Custom Bond, TIN Certificate, Income tax, Boiler use certificate, Generator Weaver, load certificate, electricity approval certificate, bond licence water test report, Fire License etc.

3.7 Training

This factory conducted different type of training on the basis of requirement. Security Awareness Training, Wastage Disposal, Environment Health and Safety, Risk Assessment Training, occupational Health & Safety Training, Anti-Corruption Awareness Training, Fire Fighter Training , PPE Training ,First Aid training, HIV Aids, Malaria, Risk Assessment training, Chemical safety & Handling Training(MSDS),Labour Law awareness, Expected Mother Awareness Training, Management And Mid-Level Management Training, Transmitted & Non Transmitted Diseases, Trauma & Serious illness, Injury Training etc. are among them.

3.8 Emergency Evacuation Plan

Each section apprehends the emergency plan in this factory as if employees and other authorities are able to follow the emergency situation easily.

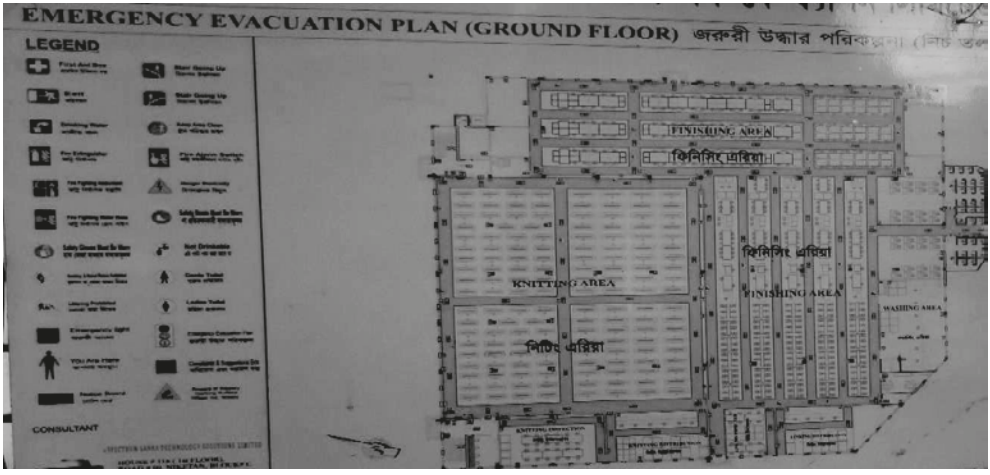


Figure 2: Emergency Evacuation Plan of the Proposed LEED GOLD Sweater Manufacturing Company

3.9 Fire Equipment and First Aid Kit

Fire equipment including Fire Hock, Fire Beater, Stretcher, Gas Mask, Belcha, and Lock Cutter, Fire Blanket, Hand Gloves, Helmet, Water Dram, Gumboot, Bucket Hose Reel, Fire Alarm Switch, Fire Alarm, Fog Light, Exit Sign, fire extinguisher (ABC, CO2) etc. are being apprehend each section with requirement. First aid kit including First Aid Box, Burna Cream, Povisep Liquid, Scissors, Surgical Gauge, Roller, Bandage, Povisep Cream etc. also being apprehend each section with requirement.

3.10 Workplace Risk Assessment

Risk assessment is a method used by managers to provide decision-making information (Through risk characterization and scoring). Separate production area apprehension of the risk assessment process depending on the area of production to be informed of the current production area. One of the section area's risk assessment procedures is given below:

3.10.1 Risk assessment

Machine Name: Swing Machine, Overlock Machine, Button Hale, Two Needle Machine, Button Machine, Weaving, Jacquard, Linking, Wash etc.

Explanation:

Possibility

1. Absolutely less
2. Chances are low
3. There are possibilities.
4. Chances are
5. More likely
6. Almost sure

Type of damage

1. Insignificant (first aid is sufficient)
2. Significant (injured but time will not be wasted)
3. Medium (casualties that will take time)
4. Large (horrific casualties)
5. Critical (mutilation)
6. Excessive damage

Risk value = probability X type of loss

,

Table 1: Risk Assessment Measurement

serial	Risk type	Risk description	Probability	Loss	Risk value = probability X type of loss	Risk prevention	Possibility after prevention	Loss After prevention	New Value
1	Component and producing product	Swing, overlock, islet the combustion of such machines and the oil produced are actually products in contact with the inadvertent fire. Fires can occur.	2	5	10	1.Fire inside the Floor is completely forbidden. 2. Initiation of welding or fire, when doing any necessary work that must work with security personnel.	2	2	4
2	Light	If there is not enough light while working on the machine. Drowsiness / headache may occur. Through this increases the tendency to get injured.	5	5	10	1. The amount of light will be kept adequate while working on the machine. 2. Regular light (Lux. Level) should be checked.	2	2	4

3	Work environment and efficiency	Pain in the waist and spine when sitting and working for a long time.	2	1	2	1.Sometimes without working one pull, Will take a break in between.	2	2	4
4	Control process	Overlock, swing, two needle machine class fire can occur in the meter if the switch is turned on by pressing with the foot.	3	5	15	1.The class of the machine cannot be switched on by pressing with the foot. First you have to switch on and increase the speed of the machine according to the rules.	2	2	4
5	Strength of different parts	If the meter, wheels, etc. of the machine are not firmly attached, they can be displaced at any time and cause an accident.	3	4	12	1.All parts of the machine needs to be firmly attached. 2. Need to do regular maintenance.	2	2	4
6	Rotating part	Machine mats and any other rotating wearables. Horrible if clothes or body parts enter. Accidents may happen.	3	4	12	1. All rotating parts should be covered with pulley. 2. Regular maintenance has to be done.	2	2	4
7	Sound	If the compressor / airline of the islet machine leaks, the noise generated can damage the ear	2	3	6	1.To maintain the airline of the machine regularly	2	2	4

8	Electricity supply	If Machine class, body, meter etc. ECC connection is not properly checked then short circuit occurred and the operator is electrocuted	3	4	12	1. ECC connect ions will be provided to all Machines. 2. ECC connections need to be maintained regularly.	2	2	4
9	Stuck with machine	If more trolleys are placed around the machine, the trolley on one side of the machine gets stuck in the other machine causing problems when getting out in an emergency.	2	5	10	1.The trolley keeping one side of the machine and the ilce mark keeping clear	2	2	4
10	Maintenance	If the machine is not maintained Regularly, it can cause problems, which can easily lead to various accidents	3	5	15	1.Regular maintenance of the machine have to check 2.Problematic machines somehow Will not be paid for work on the floor	2	24	
11	Personal safety equipment	Incidents of physical illness and injury Can happen. If the needle guard, eye guard and other safety equipment of the machine are not used while working.	3	5	15	All PPEs should be used where necessary.	2	24	

3.11 Waste Management

Garments jute, polythene bags, carton box, cone box and medical wastage are the main wastage of this factory. Jerin Enterprise collects this solid waste from this factory per week and Aichi Medical College & Hospital collects medical waste from them. They have permits for onsite waste disposal as per applicable legal requirements. Wastes are segregated by type and hazard and have maintain waste inventory and tracking system. Training is provided for waste identification, labelling and disposal.

3.12 Wastewater and Effluents Management

Washing effluents and boiler discharge are the key source of this factory's wastewater and are discharged without treatment directly to the Turag River. They have no current ETP to handle these effluents and wastewater monitoring is not carried out as per applicable requirements. But with regards to wastewater discharge, the corrective action plan has found that they planned to develop ETP, and their building is taking place.

3.13 Air Emission Management

According to the Department of Environment (DOE) and EPA, stack air quality parameter assessment of generator and boiler room is conducted by green tech environmental company. All tests were performed using continuous test methods for stack emission monitoring. KANE940 continuous analyser was used to track CO, SO_2, CO_2, NO, NOx and SPM. CO, NO and SO_2 are measured using electrochemical sensors and CO_2 is measured using the nondispersive infra-red process. In these sections, suspended particulate matter, SO_2 and NOx are found in all areas within the DOE and World Bank limit. PM 10, PM 2.5, CO, CO_2, and has not yet been set.

3.14 Water and Food Management

Groundwater and local water supply are used for factory and drinking purposes and have sufficient water use permits as per applicable legal requirements. Drinking water quality evaluation is performed by the ICCDRB Environmental Microbiology Laboratory. The total coliform, faecal coliform, faecal streptococci count, and water pH are an acceptable limit recommended for drinking water by the Bangladesh Standard and WHO.

The total aerobic bacterial count < 500/ml indicated that according to the Environmental Protection Agency, proper hygienic practice was preserved. Water usage inventory log is kept, and water use is tracked and recorded. The state of the facility including rest room, canteen room is clean. Employees carry food from their home and take it during interval. There is no food service available in this factory.

3.15 Energy use and Greenhouse Gas Management

The GHG inventory estimates the GHG emission as CO_2 emitted by JFK Fashion Ltd from its utility sector. The annual GHG emission is estimated is following scopes:

Scope 1: The direct emission from stationary combustion sources is:

Name of unit	Types of fuel used
1.Generator	1.Diesel
2.Electricity	2.Purchased electricity

Energy consumption data for direct emission in case of Diesel:

CO_2 emission	Diesel consumption of 6970 litres/year CO_2 emission is 18.716 ton/year
CH_4 emission	CH_4 emission is 0.7552692 kg CH_4/year CH_4 emission in equivalent carbon-dioxide using GWP 21 is e = 0.0007552692 ton CO_2/year
N_2O emission	N_2O emission is 0.15105384 kg N_2O/year N_2O emission in equivalent carbon- dioxide using GWP 310 is e = 0.00015105384 ton CO_2/year
Total CO_2 emission	18.716 ton CO_2/year

Scope 2: Indirect emission from purchased electricity indirect emission from electricity consumption:

Total CO_2 emission = 0.605859ton CO_2/MWH per year

Total GHG emission = 624.575 ton CO_2/year

Scope 3: Corporate value chain

Total CO_2 emission is considered as the main and major source of their GHG emission for the utility sector. In this inventory scope 3 is not considered as no data has been obtained from the supply chain of the factory. In this inventory 3 greenhouse gases emissions are considered, and these are CO_2, CH_4 and N_2O.The total GHG emission is considered for total natural gas in this factory and total electricity used by the factory in last on year. The GHG inventory has prepared in accordance with "2006 IPCC guidelines for National greenhouse gas inventories for stationary combustion". From the GHG inventory has been observed that about 97% CO_2 is emitted by indirect emission, and CO_2 emission from indirect emission 3% from diesel is negligible. Per month Energy Consumption of electricity, diesel, water and coal are given below:

Table 2: Energy Consumption

SL	Month	Electricity (KWH)	Diesel (Lt)	Water (Lt)	Coal (Kg)
1	Jan-20	42455	200	40000	19958
2	Feb-20	44400	1020	41000	27674
3	Mar-20	57600	600	40500	28925
4	Apr-20	12127	0	0	0
5	May-20	26400	0	45800	20750
6	Jun-20	120100	600	47000	30080
7	Jul-20	148800	1200	46200	33080
8	Aug-20	132000	800	45020	34920
9	Sep-20	158400	1400	42500	36911
10	Oct-20		200		
11	Nov-20				
12	Dec-20				
	Total	742282	6020	348020	232298

3.15 Land Use and Biodiversity

Initial environmental assessment has been carried out and submitted to the requisite public authority and has sufficient land use/construction permit in compliance with applicable legal requirements. Factory operations have no adverse effect on local area and biodiversity.

3.16 Hazardous Substance and Management

Hazardous substances mean chemicals that are used in the washing section and are permitted to manage or store hazardous substances in compliance with relevant legal requirements. Each chemical has its own separate dispenser and empty chemical, and hazardous material containers are immediately removed from work areas, stored in a safe designated area and regularly removed from the facility and, when not in use, bulk chemical containers are closed. They have eye-wash stations and shower stations and PPE provided to employees in the entire factory. The necessary safety measures have been provided, training is monitored by the individual concerned and all of these activities are monitored by workers at the management level. Material safety data sheet, stock chemical log and daily consumption are maintained. The inventory of hazardous materials is managed and there has adequate ventilation in the chemical storage room.

3.17 Noise Pollution

Green tech environmental company performs noise level quality assessment based on electric condenser microphone. The instrument is based on type 2 economic type. For each section of the facility, assessment was carried out to assess the sound level. In most sections, the workplace noise level was found to be adequate relative to the standard of ambient noise. But both the noise level requirement of OSHA, ECR 97 and WHO were met by the generator room and boiler room. Earmuffs and earplugs are used by high-range noise level section factory staff.

4. Conclusion and Recommendation

The clothing and textile industries have been the single dominant exporting industry since the Bangladesh Liberation War and are the major export items in terms of currency earnings. As a result of the textile industry, the country's economic growth has been remarkably steady over the last decade. The country exports 60% of its clothing products to Europe and 40% of its clothing products to America. The proposed LEED GOLD sweater manufacturing company goods have already been exported to Europe and America and are very promising for this field. While they are proposed as a LEED GOLD certified factory and called a green factory, they have some form of deficiencies in their environmental management system that have been discovered through analysis. Analysis is being carried out on the basis of a checklist for environmental compliance. For this factory, the following recommendations are given based on analysis and results:

a) Enforcement in Policy and Procedure: The Separate Production Area Policy and Procedure should be developed and enforced, and system legislation on Separate Production should be prepared and implemented.

b) Increase Service Facility: For general environmental management, current and future legislation, and a cleaner production strategy, including tracking, benchmarking and environmental performance metrics, specialized resources such as training, counselling, audits, etc. should be provided.

c) Developed Wastewater Effluents Treatment Plant: This factory has no ETP and effluents are released into water without being treated, which can have a detrimental effect on the aquatic environment. Industries should develop and regularly run usable ETPs. No water without adequate treatment should be discharged.

d) Ensure Health and Safety Management: All employees need daily health check-ups by the doctor employed by the company. Ensure that there are proper fire protection equipment and fire doors in each factory and must periodically check electrical equipment, outlets, wires and dust-free clean electrical outlets. They should make own inspection group to check the safety concerns and risk factors on a regular basis.

e) Raising Awareness: Awareness-raising, capacity-building developments and activities should be carried out among all the employee of this factory, beginning with public agencies, in relation to the environmental concerns of this factory.

Furthermore, the research findings and recommendation would be of a great use to the JFK Fashion Ltd. Authority including HR & Environmental compliance management, MS management for taking action in future research and development and also in all decision making on matters pertaining to environmental management system. Thus, it could help to create a sustainable environment.

References

Aboulnag, I., 1998. Integrating Quality and Environmental Management as Competitive Business Strategy for the 21st Century. Environmental Management and Health, pp 9-2, 65-71.

Clements, R. B., 1996. Complete Guide to ISO 14000. Upper Saddle River, New Jersey: Prentice Hall.

GoB., 1992. Environmental Policy 1992. Government of Bangladesh, Ministry of Environment and Forests, Dhaka.

GoB., 1992. National Environmental Management Action Plan (NEMAP) 1995, Government of Bangladesh, Ministry of Environment and Forests, Dhaka.

GoB., 1995) The Bangladesh Environmental Conservation Act 1995, Government of Bangladesh 1995, Gazette Notification.

GoB., 1997. The Environmental Conservation Rules, 1997 Government of Bangladesh 1997, Gazette Notification.

GoB., 2010. Industrial Policy 2010. Government of Bangladesh, Ministry of Industries, Dhaka.

Hewitt, G. and Gary, R., 1998. ISO 14001 EMS Implementation Handbook, Butterworth Heinemann Ltd, Oxford, U.K.

ISO., 1996. ISO 14001 EMS- Specification with Guidance for Use International Organization for Standardization, Geneva.

ISO 14001.,2004. Environmental Certification Step by Step, A.J. Edwards, Revised first edition, Elsevier. Key Concepts or Principles of Environmental Management Systems (EMS), EPA

Khanna, M., and. Anton, W. R. Q., 2002. What is Driving Corporate Environmentalism Opportunity or Threat? Corporate Environmental Strategy, Vol. 9, No. 4

Munasinghe, M., 1999.Is Environmental Degradation an Inevitable Consequence of Economic Growth? Tunnelling through the Environmental Kuznet's Curve. Ecological Economics, Vol. 29(1), pp89-109.

Stapleton, P., Glover, M., and Davis, S., 2001. Environmental Management Systems: An Implementation Guide for Small and Medium Sized Organizations, (2nd ed.) NSF International: U.S.A.

Left Blank Intentionally

# Equal Contribution, Corresponding Author Towards Digital Twin in Aerospace Industry

Sobhana Jahan*[1] Md. Rawnak Saif Adib*[2] Kazi Abu Taher[3]

Md. Sazzadur Rahman[4]

Abstract

A digital twin is a virtual counterpart of any physical system created by a computer. The aerospace industry is one of the vividly growing fields of this era. Along with all the possibilities, this field also comes with various risks, dangers, and uncertainties because of the usage of complex mechanisms. By appending the digital twin in the aerospace industry, we can overcome those dangers and uncertainties. As this is an ongoing field, researchers are eager to work in this arena. Furthermore, in the aerospace business, the digital twin can boost efficiency and company value. Even industrialists are also willing to get benefitted from the digital twin. That is why this paper comes up with 31 research papers on digital twin models which are used in the aerospace industry in the design, manufacturing, regular monitoring, and management phases. This paper provides several findings along with possible recommendations.

Keywords: Digital Twin (DT), Artificial Intelligence (AI), Fatigue Life, Internet of Things (IoT), Machine Learning (ML), Model-based Design.

1.1 Introduction

The digital twin (DT) is a computerized counterpart of a physically complicated system that uses real-time data to foresee results and calibrates operational adjustments. Multi-physics modeling and data-driven analytics combine to create the digital twin. Sensor and quantification systems, simulation, industrial Internet of Things (IoT), machine learning (ML), physics-based model construction, and artificial intelligence (AI) are all part of its ecosystem, which is formed using super

[1] Department of Information and Communication Technology, Bangladesh University of Professionals, Dhaka, Bangladesh, Email: sobhana@bup.edu.bd
[2] Department of Computer Science and Engineering, International University of Business Agriculture and Technology, Dhaka, Bangladesh, Email: saifadib.cse@iubat.edu
[3] Department of Information and Communication Technology, Bangladesh University of Professionals, Dhaka, Bangladesh, Email: sobhana@bup.edu.bd
[4] Institute of Information Technology, Jahangirnagar University, Savar, Dhaka, Bangladesh, Email: sazzad@juniv.edu
* Equal Contribution

high simulation software (Bäßler et al., 2020). Designing, building, evaluating, and maintaining aircraft, rockets, missiles, and spacecraft are all part of the aerospace industry. Many aerospace companies are turning to DTs to achieve their goal of lowering unplanned motor, engine, and many other system downtimes. As, the current weather conditions, asset quality, and various relatable factors are considered by DT, that is why aerospace companies love to engage DT along with the main system.

Nowadays, DT models are proven to be cost-effective, and it allows airlines to keep in service for a longer period (Gavin, 1982)(Liu et al., 2018). “By utilizing digital twin asset creation, Boeing has been capable of achieving up to a 40% refinement in first-time performance of the parts and systems used to construct commercial and defense aircraft.” - Aviation Today reports (Boeing CEO, 2018). Furthermore, DTs can manage changes in the machine and extend the machine’s average lifespan (Aydemir et al., 2020). A DT allows you to anticipate the asset's remaining life with a high degree of accuracy. For example, sensors (brake temperature, hydraulic pressure) are installed on common failure points in a DT of the landing gear. This guarantees that the DT receives data from such points on a real-time basis. This information aids in predicting the likelihood of an early failure as well as determining the landing gear’s residual life cycle.

(Tuegel et al., 2012) reviewed several new structural modeling concepts for creating and maintaining airframes using DT. Several technical hurdles like initial conditions, continuous flight loads, selecting and integrating the sub-models, and managing uncertainties in constructing the Airframe DT were also discussed by the authors. (Aydemir et al., 2020)presented a thorough examination of the available approaches and technologies, as well as the challenges confronting Digital Twin and the future of DT for aircraft.

(Glaessgen et al., 2012) investigated various DT paradigms for upcoming NASA and US military vehicles. The authors also stated that future US Air Force vehicles and NASA must be more long-lasting. Statistical distributions of the properties of materials, heuristic design, and onsite testing are all used in existing practices. Those techniques are almost certain to fall short of meeting stringent requirements. That is why the authors concluded that a significant paradigm shift is required to address the shortcomings of traditional techniques.(Phanden et al., 2021) reviewed numerous recent contributions on DT-based virtual environment modeling in manufacturing, aerospace, and automation. The purpose of this review was to lay the conceptual framework for the development of DT virtual environment methodologies and techniques.

To make it simple to distinguish DT from other related terminology like "product avatar," "digital thread," "digital model," and "digital shadow," this work (Singh et al., 2021) aims to compile the many types of DT and interpretations of DT from the literature. To understand the value that DT may provide for specific industries; the article examines the concept from its origin to its anticipated future. They have observed the origin of DT in the aerospace industry which in the future was seen to have been implemented in other sectors as well.

Table-1: Comparison between existing review works.

Ref.	Main Contribution
(Tuegel et al., 2012)	Reviewed multiple new structural modeling concepts in the airframe DT for establishing and maintaining various airframes.
(Glaessgen et al., 2012)	Investigated various DT paradigms for upcoming NASA and US military vehicles.
(Phanden et al., 2021)	Examined several DT-based modeling approaches used in aerospace (production) and automation.
(Singh et al., 2021)	Compiled a variety of DT types and interpreted all of them.
(Aydemir et al., 2020)	Provided a comprehensive analysis of the methods and technologies available, as well as the difficulties facing DT and its prospects for use in aircraft.

Table 1 provides a clear comparison between existing review works. In this table, (Tuegel et al., 2012) stated various modeling concepts on airframe DT, (Phanden et al., 2021) discussed a few DT-based simulation methods, and (Glaessgen et al., 2012) stated DT paradigm for upcoming NASA and US military vehicles. (Singh et al. 2021) presented an interpretation of a variety of DT types. (Aydemir et al., 2020) provided an analysis of various DT methods. From this table, we can say that there is no existing work that reviews DT models from design and manufacturing to continuous monitoring and management in the aerospace industry.

The block diagram in Figure 1 shows the overall flow of our paper. Note that a few subpoints in this block diagram won't match the actual subpoints of the paper. It is important to understand the structure of the paper and for that few extra subpoints are added to the block diagram. Our paper starts with an introduction

followed by the background and the research areas of DT. Then we focused on our core topic which is DT in the Aerospace industry. We have explained all the phases and sectors of aerospace and what role DT plays in each of those phases and sectors. Next, we have described the previous works on model-based design and simulation techniques. Finally, we have mentioned our findings and recommendations and concluded our paper.

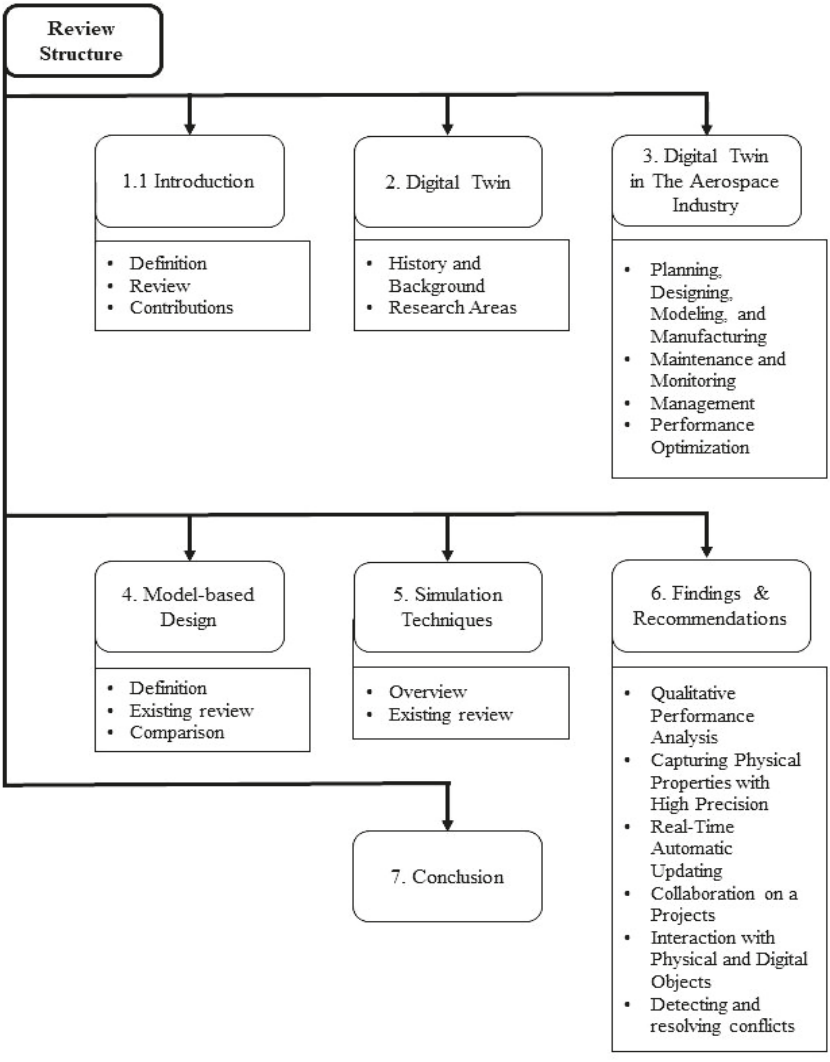


Figure 1: Block diagram mentioning the flow of this paper.

The demand for DT is growing day by day because it can assist the aerospace industry by designing, manufacturing, predicting risks, real-time monitoring, and providing a cost-effective solution. Considering these benefits of DT, researchers and manufacturers are willing to work on DT in the aerospace industry. To the best of our knowledge, there is no existing review paper that reviewed DT models used in all the phases (design, manufacturing, monitoring, management, and performance optimization) of the aerospace industry. That is why we are presenting this paper focusing on this research gap.

The main contributions of this paper are listed below:

a. Discussing the role of DT in all phases of the Aerospace industry (Design, manufacturing, maintenance, monitoring, management, and performance optimization).
b. Analyzing different model-based design approaches.
c. Discussing various simulation techniques.
d. Describing several findings & and recommendations.

The rest of the paper is structured as follows: Section two describes the basic concept of DT along with various research areas. The background of the aerospace industry, market statistics, all the stages and sectors of aerospace, and performance optimization are presented in section three. Section four comes with model-based design models that are used in this particular field. Section five consists of different simulation techniques that are used in DT. Section six comes with our findings and recommendations that include performance analysis and resolving conflicts. Finally, Section Seven provides concluding remarks. Figure 1 shows the organization of this work.

2. Digital Twin

2.1 History and Background of Digital Twin

DT is a perfect virtual representation of physical devices and systems. “Mirror Worlds” by David Gelernter, published in 1991, foreshadowed the concept of DTs (Gelernter, 1991). In 2002, Michael Grieves of the Florida Institute of Technology is acknowledged in academia and industry for creating DT as the theoretical framework of product lifecycle management (PLM) (Grieves, 2019). A famous quote from (Grieves et al., 2017) and(Greengard) who is the concept’s creator of DT is, “We have reached a point where it’s possible to have all the information embedded in a physical object reside within a digital representation.” In 2010 Roadmap Report, NASA’s John Vickers (NASA) dubbed this concept the “Digital Twin.”

The authors have offered an overview of the state of DT advancement both domestically and internationally from the perspectives of the DT idea, research advancement, interpretation, difficult issues, and DT system building (J. Wu et al., 2020). In this paper, they haven't focused specifically on the aerospace sector, but rather have provided a generalized idea of DT in addition to the above-mentioned perspectives. The physical object, the virtual or digital object, and the connections between the previous two products make up the digital twin concept. The linkages between the digital object and the physical object are data or information that flows from the real or physical object to the digital object and available information from the digitized representation to the practical environment. Later, the notion of DT was divided into other categories.

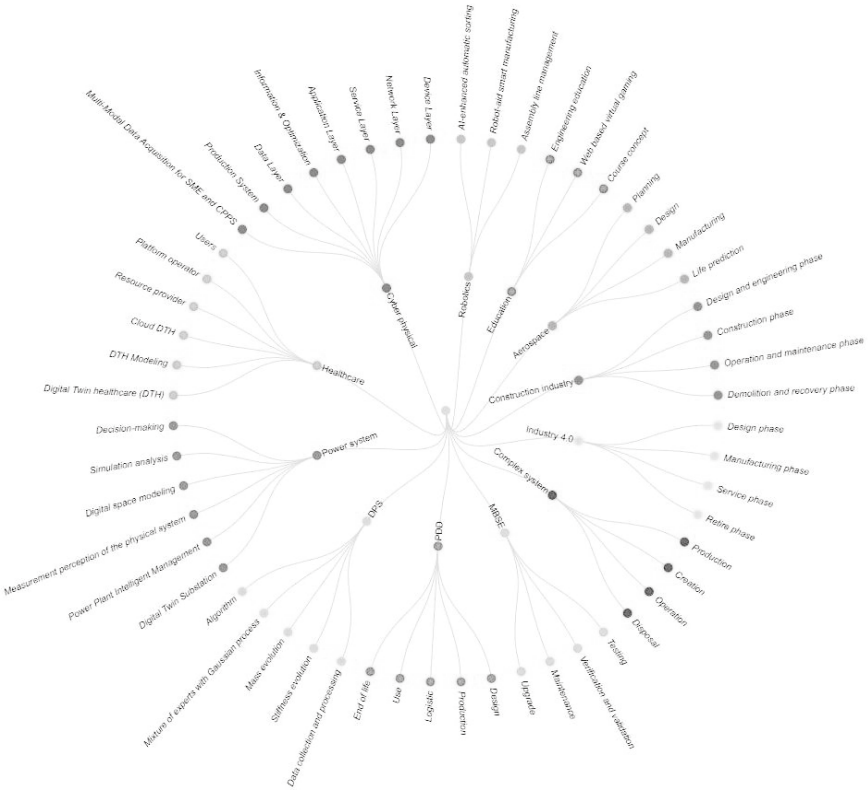


Figure 2: Research Areas of Digital Twin (MBSE = Model-based system engineering, PDD = Product design and development, DPS = Dynamical and Physical systems)

The DT Instance, DT Prototype, and DT Aggregate are the three types. The DTP entails the creation of physical products through analyses, designs, and processes. Before there is a physical object, there will be a DTP. The DTA is a collection of DTIs whose information can be used to probe cross-examination of the physical product and predict outcomes.

DTs have the potential to significantly increase the data-driven decision-making mechanisms in businesses. They communicate with their physical equivalents at the edge, and organizations use DTs to recognize the condition of physical assets, respond to changes, improve efficiency, and provide value to system applications. The next subsection is coming with the various research areas of DT.

2.2 Research Areas

A circular dendrogram is shown in Figure 2 that represents 12 different research areas of DT which are currently running in full swing. Rudimentary knowledge about these fields and in which subsectors DT is being applied are given below.

Figure 2 shows multiple sectors consisting of the implementation of DT. These are briefly discussed below. Here we have mentioned a total of 12 sectors including Aerospace. Apart from it, the other 11 sectors with the application of DT in the specific phases of each sector are listed below.

i. Construction Industry
    a. Design and engineering phase
    b. Construction phase
    c. Operation and maintenance phase
    d. Demolition and recovery phase
ii. Industry 4.0
    a. Design phase
    b. Manufacturing phase
    c. Service phase
    d. Retire phase
iii. Power system
    a. Digital Twin Substation
    b. Power Plant Intelligent Management
    c. Measurement perception of the physical system
    d. Digital space modeling
    e. Simulation analysis
    f. Decision-making
iv. Product design and development
    a. Design
    b. Production
    c. Logistic

   d. Use
   e. End of life

v. Robotics
   a. AI-enhanced automatic sorting
   b. Robot-aid smart manufacturing
   c. Assembly line management

vi. Complex systems
   a. Production
   b. Creation
   c. Operation
   d. Disposal

vii. Dynamical and physical systems with multiple time scales
   a. Data collection and processing
   b. Stiffness evolution
   c. Mass evolution
   d. A mixture of experts with the Gaussian process
   e. Algorithm

viii. Model-based system engineering
   a. Testing
   b. Verification and validation
   c. Maintenance
   d. Upgrade

ix. Cyber-physical
   a. Multi-Modal Data Acquisition for SME and CPPS
   b. Production System
   c. Data Layer
   d. Information & Optimization
   e. Application Layer
   f. Service Layer
   g. Network Layer
   h. Device Layer

x. Medical and healthcare
   a. Digital Twin Healthcare (DTH)
   b. DTH Modeling
   c. Cloud DTH
   d. Resource provider
   e. Platform operator
   f. Users

xi. Education
   a. Engineering education
   b. Web-based virtual gaming
   c. Course concept

3. Digital Twin in The Aerospace Industry

The aerospace industry is a worldwide conglomerate of governments and individuals dedicated to product development, research, and manufacture in the areas of flight, atmospherics, and beyond which began in the United States in December 1903, after Wilbur and Orville Wright exhibited controlled sustained flights. It is among the most important and strong sectors in the United States. Various applications of DT models in the aerospace industry are described in this section.

To dispel myths that prevent the sustainable use of digital twin technology for protection systems, this article explains the complexities of innovation for the aerospace sector particularly and others generally (cnc et al., 2021). It includes a thorough analysis of digital twins and the parts that make them up. In their paper, they have described the concept of DT, its major components, some industrial applications, and its importance in the aerospace industry. They also mentioned the business challenges, modeling, infrastructure, visualization, and future trends.

3.1 Planning, Designing, Modeling, and Manufacturing of Aircraft

During the planning and designing phase, DTs can be used as virtual prototypes that can be modified to try new simulations. It can also be designed before investing it in a strong prototype. For modeling and manufacturing any complex part (e.g., the engine and the black box) of the aircraft, it is truly beneficial to make a virtual version of it using DT and validate it before building the physical one. This reduces the number of repetitions required to get the item into production, saving time and money.

3.1.1 Planning Stage

A closed-looped air cargo load planning system based on DT enhances container loading operations constantly using a virtual environment (Closed-loop, 2022). This work's major goal is to improve loading strategies, eliminate effort in container warehousing, and provide a feedback system for appropriate loading scenarios. The author's proposed model is constructed using three components and those are optimization simulation, physical operations systems, and Virtual Reality (VR) systems. Here, the optimization module contains the decision support tool which helps to create the simulation of cargo pallets in an aircraft. Based on the cargo nature, contour, weight, size, destination, and dangerous goods classification, the cargo pallets are simulated. This simulated data serves as source information and is transmitted to VR devices via a cloud environment. The VR module is a CAVE-based system with a capsule construction and four 3D projections and displays. This VR system detects the user's current location with

sensors and then projects dynamic scenarios with an area of vision and stereo presentation. The simulated data from the optimization module is obtained and imitated by the VR module to verify the effectiveness of stacking the cargo onto the aircraft.

If overstuffed cargo is to be packed, the module can inspect that before loading it to the aircraft door. Users can engage with the aircraft stacking slots, and cargo can view the simulated actions and discover anomalies that require recovery.

Table-2: Comparison between various DT models used in the Aerospace Industry for Design and Manufacturing.

Ref.	Focused Area	Main Contribution	System Architecture	Limitation
(Closed loop, 2022).	Cargo load planning	A closed-looped air cargo load planning system that uses DT to simulate cargo loading operations to improve cargo loading operations.	A three-module-oriented DT model and those three modules are optimization simulation, physical operations system, and VR system.	The capabilities of this model for different aircraft designs, such as one and two-row aircraft, and primary and bottom decks, should be studied, analyzed, and expanded.
(Chang et al., 2021)	Designing aircraft assembly line	DT was able to create an architecture for AAL to improve its efficacy, quality, and visibility.	An interactive architectural knowledge repository is created, and design defects are identified based on its contents.	The proposed architecture's effectiveness and feasibility were not compared to other similar standard architectures.
(Liu et al., 2021)	Machine modeling	A biomimicry-based modeling method that can	The geometry module, behavior	The speed and accuracy of the

		provide high-fidelity multidimensional DT in the field of machining.	module, and context module comprise a three-module DT system.	model should be increased.
(Jinsong et al., 2019)	A modeling approach for manufacturing	Modeling and operational approach for the DT in the area of aerospace manufacturing.	The proposed DT models had 3 sub-models-product DT, process DT, and operation DT.	The authors should carry out workshop-oriented modeling and operation using DT technology.
(Zambal et al., 2018)	Creating a manufacturing dataset	A manufacturing dataset is created using real-time sensor and machine data to fabricate a portion of an aircraft wing cover (lower).	Sensor data was gathered using Gigabit Ethernet, while device information was analyzed using PowerLink. All of this data was pooled in two separate PCs capable of low-level computation. Data on defects was transmitted to two distinct servers. Part verification is performed by one of the PCs. On the	Because many unexpected scenarios emerge throughout production, requiring immediate decision changes depending on the current situation, the decision-making process should have been incorporated.

			other PC, the manufacturing database was hosted.	
(Mandolla et al., 2019)	Additive manufacturing	A case study for the aerospace industry where a DT-based end-to-end additive manufacturing model is described.	A layered model for creating a DT manufacturing model. The four layers are design, information, creating a blockchain, and hashing.	Only four layers are described. The testing and validation phases were not described properly.

Physical operations systems can cover the operations of aircraft cargo loading and the gate-in terminal on the air side of the airport. When the cargo is analyzed at the terminal gate, sensors inside the entrance terminal are employed to calculate the exact cargo dimensions. It detects cargo shape and weight and transmits this data to optimization and virtual reality systems via cloud storage. When enormous cargo is loaded aboard the airplane via the primary deck ground rollers, loading platform, and container doors, sensors are employed to measure any irregularities in activities and enhance the modeling procedures. One limitation of this work is the model's capability for various aircraft configurations was not considered here.

(A. Hänel et al., 2020) uses the instance of components in the aerospace sector to explain a technique for developing a DT depending on planning and process data for machining operations. They have provided descriptions of the information methods, fundamental composition, and pertinent calculation methods for identifying pertinent process features and its implementation will be represented in a milling process for the aerospace industry. Their structure for the machining process consists of the following steps-

i. Data acquisition
ii. Processing the acquired data with various models
iii. Data calculation
iv. Application of the calculated data

In the first process, they acquired 5 types of data. These are workpiece data, process data, technology data, machine data, and tools data. These data were further processed using 5 processed models namely, kinematic models, tool engagement models, path inaccuracies models, cutting force models, and surface quality models. And finally, before application, they visualized the model in software to check all the processed parameters. This technology makes it feasible to create a digital process twin to a milling process for the manufacture of aerospace elements with a significant number of pertinent process features.

3.1.2 Designing Stage

Digital-physical engagement and self-evaluation are two capabilities that modern aircraft assembly lines (AALs) must have. To provide this facility, (Chang et al., 2021) proposed a design architecture of knowledge-enabled DT for AAL to improve AAL's effectiveness, quality, and visibility. The proposed model is designed to introduce DT and knowledge. To begin, a one-of-a-kind vibrant layout knowledge repository is built to accommodate the multiple DT operations that occur during the design process. Ultimately, knowledge-based intelligent assessment is being used to comprehend and discover architectural defects that might lead to the layout plan's subsequent improvement. One limitation of this research work is the effectiveness and feasibility of the proposed architecture were not evaluated with other similar standard architecture.

In the designing phase, we also need to think about the design of the machines that will be used for the aircraft. Because this concern will impact the manufacturing phase. In general aspects, multiple high-performance machines are being used in the aerospace sector. Most of them fall under CNC (Computer Numerical Control) Machining. Certain factors are kept in mind when CNC machines are made. These are Lightweight metal components, CNC rapid prototyping, Complex design fabrication, AI-aided CNC machining, and High-grade precision. All the machines that are manufactured must meet these requirements.

3.1.3 Modeling Stage

To remove the scarcity of effective modeling methods in the aerospace industry, (Liu et al., 2021) proposed a biomimicry-based modeling method that can provide high-fidelity multidimensional DT in the field of machining. For this purpose, the authors have proposed three modules that are the geometry module, behavior module, and context module. These three modules are interactive and can imitate the real physical model. In the geometrical module, the manufacturing features are analyzed to understand the machining characteristics. The system gets the actual information with the help of performance-measuring and size-measuring

instruments. As this module enables self-updating so the module updates its information systematically. With the help of the fusion technique, the real and the virtual system are merged. In the behavioral module, with the help of a system database, a theoretical process model and material information are obtained. Computer-aided Engineering (CAE) software help to obtain the structural features. The offline feature technique aids in the accumulation of physical attributes for the creation of the physical module before manufacturing. The rest of the physical property information is obtained and updated in real time by the digital twin system using surface measurement.

The establishment of the DT-based context mimic model includes two steps which are machining sequence generation and the machining process information collection. The development of the manufacturing series intends to build the operational route of the manufacturing entity in the procedure organizing before machining. During the computation, the system gathers pertinent architectural, temperature, current, motion, voltage, and actual procedure statistics. The modifications made to the computation product are then recorded by the data-capturing mechanism. The information obtained is examined and utilized to generate that product's processing information in the final stage.

In the area of production, (Jinsong et al., 2019) proposed DT models which had three sub-models and those are product DT, process DT, and operation DT. To begin, across the product designing and production stages, a unified model was utilized to develop the product DT, letting information be exchanged and distributed throughout phases. By adding interface and control qualities, process DT supported the interconnectedness and convergence of digital and real environments inside the manufacturing system.

Secondly, the multivariate model fusion was created by the authors for constructing the operation DT metamodel. Various interactive behaviors could be designed and simulated using the flexible mixture and integration of different operation DT metamodels.

Finally, the authors constructed a systemic component manufacturing cell AutomationML framework. It was a revolutionary method of carrying out activities in the manufacturing context involving products, techniques, and resources. One limitation of this work is the authors should carry out more detailed work on the workshop-oriented modeling and operation using DT technology.

3.1.4 Manufacturing Stage

Manufacturing a part of aircraft using carbon fiber composite material is complex work. DT can be a helpful solution to this type of manufacturing. With the help of

a manufacturing database, creating this DT model will be very easy. To do so, (Zambal et al., 2018) proposed an architecture that collects different sensor data and machine data to create a DT that will help in manufacturing the lower wing of an aircraft. Sensor data was acquired over Gigabit Ethernet, whereas machine data was collected via a PowerLink interface in this case. All of this information was pooled in two separate PCs that do low-level computation. Fault data was routed to two different servers. One of the PCs performs a part verification. The manufacturing database was hosted on the other PC. In terms of information, the Gigabit network was used to transfer the most data. Here, the HDF5 file format was used by the database, which had a flexible structure that allows for the storage of multiple types of data.

The simulation program was used to poll occurrences regularly, and it would start analyzing once it had been informed that the faulty data was sufficient. A limit of security was estimated after the preliminary analysis through the mathematical framework. In the latest version, in response to a user request, a thorough bounded calculation was carried out. The database had been updated with the simulation findings, and the characteristics endpoint had been updated with the limit of security. The selection assistance tool came next, which updated the user with the most recent knowledge of the production cycle by checking the events per 15 seconds. With the help of Blockchain, (Mandolla et al., 2019) presented a case study for the aerospace industry where a DT-based end-to-end additive manufacturing model is described. For creating this model, the authors have constructed four steps which are design, block creation, information, and hash verification. With the help of "CatiaV" (3D CAD designing software), the authors have designed a blade compressor for airplanes. The information was infused into the common software named "Ultimaker Cura" by utilizing the design created in CatiaV5. For creating the blockchain, public and private were developed. The authors had altered the original basic scheme to create a Blockchain. In the initial subsection, the authors fixed every constant and primary function. In the final section of the code, the "readFileSync" function was inserted into the block. After all of these steps, the hashing operation took place. One limitation of this case study is the testing and validation step was not described properly. **Table 2** summarizes DT models.

3.2 Maintenance and Monitoring of Aircraft

The modern aviation sector is transitioning from responsive to aggressive and predictive service to increase platform functional reliability and effectiveness, expand its useful life span, and reduce its lifespan costs. In aviation service, the DT is used to reflect both the aircraft and the MRO (Maintenance, Repair, and

Overhaul) activities. For the progress, delivery, and procedure of an air vehicle, a well-qualified performance model for an airplane arrangement is required. (Kraft et al., 2018) discussed the methods and characteristics required to create and implement a single authoritative truth reference as an integral component of a DT strategy for aerial vehicle acquisition and maintenance.

3.2.1 Fatigue life prediction

The number of stress cycles of a particular character that an object can sustain before failure is defined as fatigue life. Submicroscopic cracks are the cause of fatigue failure. These appear on a material's surface. Those cracks eventually come together to form a visible crack. Crack initiation, Total life, and crack growth are the three main methods often used to predict fatigue life.

Using DT to predict the duration of aircraft components and guarantee their structural integrity, (Tuegel et al., 2012) established a theoretical foundation. To begin, a series of missions must be allocated to a specific aircraft. It is formed as a reasonable approximation of the flying trajectory and maneuverability that will be traveled during the quest. The aircraft's computational fluid dynamics (CFD) method is then "flown" virtually through all those aircraft to approximate the heavy loads and surroundings the aircraft will encounter. Aerodynamic forces on the airframe are implemented to the functional Digital Twin finite element model (FEM) over the interval of time of the flight as the airplane is "flown." The FEM and CFD models are tightly coupled, allowing the impact of aero-elastic vibrations and functional deflections on the aerodynamic flow to be studied. The temperature, stress, and vibration predictors data are sent to the damage predictor with the help of the driver, and the damaged states are estimated. Through this, the structural system reliability assessment phase takes place. Finally, after the execution of data, the aircraft's health information is passed to the digital twin to refine the information, and those refined data are passed to the reliability assessment block for damage prediction. Because it is a conceptual model, some new issues may arise during implementation.

(Leser et al., 2020) developed a systematic way to decrease unpredictability in fatigue life predictions by combining in-situ diagnostic tests in a probabilistic manner. As more of this diagnosis and treatment is achieved using Monte Carlo methods, to finally create probabilistic forecasts of fracture state throughout the duration of a complicated geometrical trial sample and anticipate residual stresses with decreasing unpredictability.

It is claimed that the suggested DT system is appropriate for exhaustion lifespan diagnosis and has to be developed considerably deeper with an emphasis on increasing implementation reality by demonstrating the potential to anticipate

accurately as well as in the face of ambiguity. The research's conclusions showed that the integration of the suggested diagnostic techniques enabled the application of DT for predicting exhaustion lifespan in a lab environment. The four probabilistic edges of life projection's average scores were all under 0.2% of the actual final result. All four deterministic predictions of the product's residual lifetime was within 10% of the genuine residual lifetime seen throughout the experiment.

(Keivanpour et al., 2022) designed a methodology for estimating the remaining useful life (RUL) of retrieved parts that includes fuzzy modeling and DT. The presented framework incorporates DTs: disassembly process, health management, and retrieval parts. A fuzzy smart decision-making algorithm produces the rules for estimating a part's RUL based on various sensor measurement techniques and operational setups. The model's implementation in aircraft engines is also discussed by the authors. For illustration, the PHM08 prognostics records database was used here. The data collection includes deterioration information from 21 sensing devices along with operational and maintenance parameters. Time series analysis's initial 14% are seen as robust, while its latter 14% are regarded as a failed system.

3.2.2 Health monitoring & maintenance

Aircraft health monitoring methods rely on Digital Twins to use real-time data captured by various sensors incorporated into aircraft parts to improve the aircraft's safety and reliability. It aids in the reduction of operational costs, support costs, and unexpected accidents. Quality deviation problems are difficult to find and fixing them requires a while. A quality variation regulation methodology for DT was put forth by (Cai et al., 2021). DT simulation based on the asset management shell approach can be used to extract and combine multiple origins and diversified quality variation information. In addition, a quality deviation scheme based on DT has been developed to address the second issue. The FP-growth association rule method analyzes the aircraft quality deviation information in this system, and the outputs are displayed through the scheme to guide the assembly facility.

The systemic health status of an aircraft differs from one to the next due to differences in production, mission history, material properties, pilot diversity, and so on. (Li et al.) built an aircraft health monitoring framework for the prediction and diagnosis of each independent aircraft using the idea of dynamic Bayesian networks (DBN) and demonstrated the proposed scheme using an aircraft's wing fatigue crack propagation example. The DBN model is used in diagnosis to monitor the progression of time-dependent factors and fine-tune time-independent

factors. In diagnosis, the DBN model is used to make probabilistic predictions about crack formation in upcoming loading time steps.

The book section of (Wang et al., 2015) includes an introduction to DT, its generation, and its different components. Then the maintenance process and the applications of DT in aircraft maintenance are described. MRO (maintenance, repair, and overhaul) is a requirement for commercial aircraft to ensure the safety of passengers which are often done with DT's assistance. Four technologies drive the development of DT. These are Big data, IoT, modern analytics, Cloud computing, Accessibility, and interaction. They gathered real-world MRO utilizing different IoT sensors aboard the aircraft as part of their system architecture. After that, the raw data is processed on a cloud or edge platform before connecting with the DT database. They also keep a close eye on what's happening in the hangar. Then the Virtual MRO is created using DT technology, and different operations are done using real-world data to match the expected outcome. The MRO facilities system offers pertinent services for managing service operations, monitoring current health, and predicting prospective states.

To prevent the major loss, early diagnosis of damage in composite structures is critical. The computer method that forecasts the dynamic features of a chopper sandwich construction struck by a bird was developed by (Ibrion et al., 2019). The prototypes are intended to serve as simulated procedures for a brand-new defect identification approach that uses a digital twin in the coming years. A high-fidelity (HF) LS-DYNA FE/SPH model was first developed and tested against soft body effect studies. After that, a low-fidelity (LF) model that was operationally efficient was developed and compared to a high-fidelity (HF) framework. It was discovered that the high-fidelity framework can correctly forecast the strains captured by the detectors, and the volume of the projected debonding region toward the sandwich panel's frontal side matches the experimentally measured delamination area extremely well. The HF framework can be viewed as a simulated serious damage discoverer and damage outgrowth predictor before the scheduled observation, meanwhile, the LF framework can be utilized as a quick numerical tool for determining the load situation.

(Singh et al., 2021) provided a comprehensive overview of DT in aircraft manufacturing, as well as a proposed model for collecting and examining information about aircraft wings. Crack growth and spreading are common issues in airframes. According to the authors, the suggested framework will allow professionals and information management experts to collaborate closely in the development and maintenance of DT. The physical level of the author's four-layered architecture recognized the system's current needs, which were recorded

in the shape of developed-in accordance. Following that, the information accusation layer comprises various on-ground sensors and IoT-enabled techniques for reading, pre-processing, and transferring data. The amount of fatigue on the wing was identified by this layer. Data were converted into useful information and insights at the model layer. A fatigue crack computation behavioral model was developed for the wing structure's design and analysis purpose. Information had been subsequently organized and saved throughout the DT lifespan at the Data modeling level, which is where it is now. The proposed framework's dependability has not been assessed. For the ongoing safe operation of an airplane, load levels should be maintained under a reduced load-carrying ability to avoid unsteady, catastrophic harm propagation throughout a flight. As a result, for constructions with complex harm/damage configurations, the inability for exact real-time forecasts of serious harm size and safe load-bearing capacity is required. As a result, (R. Seshadri et al.) proposed a method for monitoring damage using guided wave reactions. The signal dampens in some orientations and tends to reflect in others when the guided wave communicates with damage. This results in changes in signaling amplitude and phase changes among structural reactions that have been fractured or cracked and those that have not. Only a genetic algorithm (GA) driven optimization technique is used to analyze the collective signal reactions at several previously selected sensor locations to produce an accurate estimate of the size, position, and orientation of the damage. The damage size and direction are obtained by minimizing the variation between the reference reactions and the reactions obtained from the finite component calculation of the propagation of waves for various typical crack formations, geometrical damage, and sizes. A Piezoelectric (PZT) sensor-based Guided wave may be a better way to monitor aircraft damage.

Table-3: Comparison between various DT Models used in the Aerospace Industry for Maintenance and Monitoring.

Ref.	Main Contribution	Limitation
(Tuegel et al., 2012)	A theoretical framework for estimating the lifespan of airplane structures and ensuring their structural integrity utilizing the DT.	Because it is a conceptual model, some new issues may arise during implementation.
(Kraft et al., 2018)	Discussed the methods and characteristics required to create and implement a single authoritative truth	-

	reference as an integral component of a DT strategy for aerial vehicle acquisition and maintenance.	
(Keivanpour et al., 2022)	DT and fuzzy modeling were used to create a system for determining the retrieved parts' residual useful lifespan (RUL).	-
(Cai et al., 2021)	A DT-related performance variation controlling approach. DT simulation based on the asset management shell approach can be used to extract and combine multi-source and diversified quality variation information.	-
(Li et al.)	An aircraft health monitoring framework for prediction and diagnosis of each independent aircraft using the idea of dynamic Bayesian networks (DBN) and demonstrates the proposed scheme using an aircraft's wing fatigue crack propagation example.	-
(Ibrion et al., 2019)	The computational system that predicts the dynamic characteristics of a helicopter sandwich structure hit by a bird.	-
(Singh et al., 2021)	A model for collecting and examining information about aircraft wings.	The proposed framework's reliability has not been evaluated.
(Seshadri et al., 2017)	A method for examining damage using guided wave reactions.	-
(Zakrajsek et al., 2017)	A DT model to improve the tire touchdown wear estimation.	Several aspects of the DT model introduced in this study remain unfinished.

		Increasing the levels and testing the additional factors are two major areas for improvement.
(Leser et al., 2020)	Developed a systematic way to decrease unpredictability in fatigue life predictions by combining in-situ diagnostic tests in a probabilistic manner.	-
(Wang et al., 2015)	Explained the maintenance process and the applications of DT in aircraft maintenance. Discussed MRO and what services it provides.	-

Underperforming airplane tires can result in high expenses as well as an increase in the aircraft's integrated logistics and environmental footprint. Airplane tire wear methods are intricate and reliant on a slew of interconnected variables. A DT model of a particular plane tire at touchdown was established by (Zakrajsek et al., 2017) and it is used to improve the above tire touchdown wear estimation. The derivation of a physics-based formula, slip Wear percentage is used in a non-linear touchdown wear reaction model. Based on the above-mentioned variables, high-fidelity testing data is detailed in this paper. The response model was created as a function of variable factors that can be seen in a field setting. After that, the DT method is used to calculate the Probability of Failure (POF) for various sink rates, yaw incidence angles, tire situations, and touchdown velocities. The findings of the DT touchdown model indicate future potential benefits for airplane mission decisions, such as cost reductions and tire continuous monitoring at touchdown. Several aspects of the DT model introduced in this study have yet to be completed. Two major areas for improvement are raising the levels and testing the additional factors. **Table 3** summarizes all the mentioned models along with their limitations.

3.3 Management

3.3.1 Airport management

The Digital Twin concept allows an airport authority to make decisions inside a data model, analyze the results, and optimize the outcome. The truth is that airports are lagging in embracing DT, which is already widely used in high-tech manufacturing and the fields of engineering, architecture, and construction.

Related to time savings, cost cutbacks, design, and performance improvement skills, these sectors already get benefitted from the immense gains in productivity.

In general, an airport DT can offer new sections of strategic value, such as:

a. Platform vision is achieved by combining existing resource operational optimization with combined stakeholder circumstance development and future strategic initiatives.
b. New Innovation and savings are made possible by DT.
c. System advancement is based on collaborations with the tech industry to simulate new operations and commercial services using DT.
d. Revenue generation for first-mover contractors who develop airport-specific applications for global distribution to other airports.

Table-4: Comparison between various DT Models used in the Aerospace Industry for Management.

Ref.	Main Contribution	Limitation
(Saifutdinov et al., 2020)	A DT model-based airport traffic control scheme based on the Decision Support Tool.	The combination of DT with ML was explained in a theoretical manner. The practical application procedure was not described.
(Reifsnider et al., 2013)	Based on the observation and scientific justification of fluctuations in several physical substance characteristics, a rational engineering approach to real-time composite structure "tail number" diagnosis.	The utilization of dielectric spectroscopy in introspective geology, sensing and danger detection, electrical circuitry, and a variety of bio-based methods is still in its infancy.
(Zaccaria et al., 2018)	A scheme for aircraft fleet monitoring, and management.	This model is not tested using real flight data.

A DT of an airport contains more than just the building infrastructure; it also includes the associated operational processes. It acts as a major data repository, allowing it to enhance performance and acquire data more quickly. Advertisement, assets management, operational processes, engineering, and servicing departments are among the business sectors that will benefit. The DT model should also encourage a climate conducive to digital transformation. A business strategy is

required for the successful implementation of such virtual models so that DTs can be used successfully by organizations to produce positive results.

(Saifutdinov et al., 2020) proposed a DT model-based airport traffic control scheme based on the Decision Support Tool. The authors described various types of relationships that can exist between a DT and an actual physical element. The innovative aspect of the investigation is the continuous linkage between the DT and the physical thing in question. The most efficient response seems to be a distinct and precise division of the information system and applications in the manner of the DT. The DT can therefore be utilized as a substantial information source for the development and evaluation of different approaches to centralized ground transportation control at airports. Theoretical explanations were provided for the combination of DT and ML. The procedure for carrying out the practical application was not described.

3.3.2 Fleet management

The goals of fleet management are usually readiness, accessibility, risk reduction, and cost reduction. Data management is required to handle massive datasets from various platforms. The science and technology behind the interpretation used in each "tail number" are, however, at the heart of any successful, assertive fleet management.

(Reifsnider et al., 2013) discussed a logical engineering scheme for real-time composite structure "tail number" diagnosis based on the measurement and scientific explanation of changes in multi-physical material characteristics. The strategy has been demonstrated with examples, and application paths will be mentioned. There will be connections made between durability, trustworthiness, threat, and liability. The use of dielectric spectroscopy in exploratory geology, sensing and threat detection, electrical circuitry, and a variety of bio-based methodologies is still in its early stages.

(Zaccaria et al., 2018) developed a scheme for aircraft fleet monitoring, and management. The framework takes a multi-level strategy: it starts with monitoring thresholds, then isolates troublesome engines, which are then subjected to a fault detection technique. Various fault isolation, recognition, and estimation methods were explained and compared by the authors. This theoretical strategy is evaluated on fleet data obtained by a turbofan engine efficiency model, considering flight-to-flight and engine-to-engine variants as well as sensor measurement uncertainties. The above-mentioned model has not been validated using real-world flight data. **Table 4** summarizes all DT models used in management along with their limitations.

3.4 Performance Optimization

With digital twins actively monitoring and duplicating the performance of real aircraft systems, operators and companies will be able to predict and prevent a much broader range of problems before they happen. As a result, uptime and airplane availability will be significantly improved.

The precision spool valve is a crucial element of the electro-hydraulic servo motor control system, and its behavior has a significant impact on automotive and aerospace flight control. (Tang et al., 2021) proposed a novel technique for evaluating the mating output of precision spool valves that uses DT technology to account for identifying surface topographical errors. First, a basic framework for analyzing highly precise spool valve mating performance using the proposed DT. Following that, key technologies such as construction interface geometrical modeling, corresponding behavior modeling, and performance comparison are investigated. Finally, a quantitative relationship is discovered between the mating parameters and the spool valve's oil leakage. The technique is put to the test in a real-world scenario. This technique can provide theoretical justification for precise estimates and examination of the spool valve's mating results. In this mating performance optimization model, geometrical topography parameters were not taken into account.

(Sisson et al., 2021) presented a methodology for optimizing rotorcraft flight variables to reduce the stress on serious mechanical parts based on their current and projected states while meeting technical constraints. To assist probabilistic diagnosis, and utilization, a digital twin strategy is being pursued. A detailed rotorcraft analysis was used to determine the rotorcraft's load capacity under specific flight circumstances, and a finite element analysis was used to anticipate the correlating stress in the element of interest. After each flight, diagnostic data is taken and then used to approximate the disparity in the damage prediction model. The up-to-date prediction technique is used to predict the component's condition for a given future mission. The ideal flight specifications for the upcoming mission are then recognized by taking into account the component's anticipated state and reducing stress in the element under ambiguity. Numerical experiments are used to demonstrate the proposed method.

A two-stroke large-fuel airplane engine with a poppet control valve has an inefficient manufacturing process since the fuel interchange system contains several components. Furthermore, due to the higher valve variables, actual experimental optimization wastes a lot of money and time. A DT-based optimization framework for the system was put forth by (Xu et al., 2021) and uses a variety of DT components to digitally recreate, enhance, and produce the

parameters while engaging with and storing the information. To rectify problems as they happen throughout the optimization process, the DT elements leverage actual real-time feedback data from production measurement techniques and system effectiveness testing. The results demonstrate that the iterative computation for the optimal parameters is successfully carried out, and the simulated engine model with response and adjustment is quite realistic and convincing when compared to test results. Real-world testing demonstrated the effectiveness of DT-based refinement, with both power and gas transfer effectiveness increasing by around 4% at various engine loads and speeds. Other than the gas exchange, this aero-engine optimization technique did not focus on any other fuel system.

The variability in results among independent systems increases as systems age because they may be conducted in various environments and with varying degrees of severity. As a result, a need has been recognized to shift paradigms and connect models to personal vehicles rather than fleets of identical vehicles. This allows for the acquisition of performance-specific characteristics and, as a result, a reduction in performance estimation uncertainty. (Jeon et al., 2019) proposed a framework for building and demonstrating a DT representation of a quadcopter to improve estimates of end-user-relevant metrics like maximum range. The performance predictive capability refers to these metrics. To enhance this capability, models for the multiple elements of the quadcopter structure are being developed. These metrics of interest are predicted by the models before undertaking experiments. The results of the research are analyzed after it is completed to gain a better realization of the physics and to achieve a good prediction accuracy of the models by upgrading them. Only one phase-off flight (hover) was validated in the framework.

4. Model-based Design

Model-based design is an arithmetical and visual approach for problem-solving in the design of critical control, data processing, and communication frameworks. It has numerous applications in gesture control, industrial machinery, aerospace, and automobiles. In the industrial domain, managing or handling the lifecycle of the components of a complicated and safety-critical process, from core conceptual design to operational support, is still a noteworthy challenge. Model-based design (MBD), model-based engineering, and other tasks involving the identification of requirement specification of the system, physical and functional attitudes, and model-based engineering is subjected to considerable trade studies aimed at choosing the most appropriate solution.

(Bachelor et al., 2020) discussed how the model-based DT and threads ideas will alter the way of process to overcome issues such as federated IT infrastructures

and services. Notably, perspectives of non-lifecycle and lifecycle inter-operability are explained, with a particular emphasis on the use of benchmarks for lifecycle help and diverse simulation. As a case study, the architecture of an ice protection system for a regional aircraft was chosen and described by the authors.

(Miller et al., 2018) tried to improve method interconnectivity by incorporating spatially connected non-geometric data into computer-aided design schemes. Inside the computer-aided design widget toolkit, a tool was developed to capture, illustrate, and search for spatial information. This allows both researchers and users to access information previously only available in separate software from within the CAD tool.

A safety-critical system's layout necessitates an accurate prediction of its RAMS (reliability, availability, maintainability, and safety). This is a tough action since the RAMS analysis tries to deal with the hazard assessment of parts of the system, which is never easy to abstract at the conceptual stage. Through functional and dysfunctional assessments, (Brusa et al., 2021) examined how model-based systems endorse this work and drive the allotment of system reliability. The toolchain must be set up before the proposed approach can be implemented. It must be agreeable with existing product creation practices, guidelines, and tools in the industrial environment.

Table-5: Comparison between various DT Models used in the Aerospace Industry for Model-based Design.

Ref.	Main Contribution	Limitation
(Bachelor et al., 2020)	Discussed how the model-based DT and threads ideas will alter the way of the process to overcome issues such as federated IT infrastructures and services.	Integration of voice recognition or virtual reality with ML-based analytics was not shown.
(Miller et al., 2018)	Tried to improve method interconnectivity by incorporating spatially connected non-geometric data into computer-aided design schemes.	This model was just a starting point for creating a digital twin using previous behavioral data from the CAD model.
(Brusa et al., 2021)	Examined how model-based systems endorse RAMS analysis	-

	and drive the allotment of system reliability.	
(Liang et al., 2020)	Introduced a real-time full-field displacement recognition method.	-

Full-field deflection perception in the high exactness production sector, such as aerospace production, is crucial. (Liang et al., 2020) presented a real-time full-field displacement recognition method for the integration of digital multicast dispersion tracking and grid achievement ideas. To begin, a theoretical full-field dispersion perception model is established based on the measured information from the multi-points. Major elements of full-field displacements are obtained by breaking them up into numerous discrete points that are both recognized and unrecognized and then creating a link between the identified spots and the full-field separations. The full-field displacement prospective model's solution is then proposed by the authors. The optimization approach is used to function the model, while the pseudo-code is used to put forward, based on big data and the matrix completion principle. Finally, full-field displacement perception experiments are carried out. Experimental studies show that the median error should be less than 0.054 mm and the maximum error of the displacements should be less than 0.094 mm. This result was very promising as it if providing high precision and efficiency for large aircraft gathering and arrangement. **Table 5** summarizes all Model-based DT.

5. Simulation Techniques

Simulation is a broad topic with a hazy definition. In general, DTs can be considered a simulation method, but not all simulation models are DTs. Regarding the degree of visualization, the strategies used to adopt DT simulation nowadays can be classified into three main levels (Ezhilarasu and Jennions, 2020). Abstract frameworks are the first level. The topology method and Simulink or Matlab model represent an abstract simulation using simple symbols and lines (West and Blackburn, 2017), (Dröder et al., 2018). Instead of displaying the shape and detailed information, the above type of model emphasizes promising principles of physics guided by the physical object (Zhidchenko et al., 2018). In today's DT applications, abstract models, particularly Simulink or Matlab models, are really common(Aivaliotis et al., 2019). The low-cost visual representation of physical items is an added benefit of the abstract model (Cimino et al., 2019). The major drawback is that users or customers without a specific level of prior knowledge will be unable to comprehend the models, particularly since they are commonly non-3D models (Rocca et al., 2020).

The study presented by (R.K. Phanden et al., 2020) provides a variety of current accomplishments on simulation-based DT and DT-based simulation modeling put forth by academics working in the fields of robotics, manufacturing, and aerospace. In the simulation of the aerospace industry, simulation techniques like Monte Carlo simulation, Finite Element Methods (FEM), and Computer-Aided Engineering (CAE) are being used to acquire the simulated data for various operations of the aircraft.

The 3D models are at the next level. 3D models are much more convenient to present the details of physical entities than abstract models. 3D modeling has progressed considerably in recent years (Breaking news). Nx, Computer-aided Design and Drafting (CAD), Analysis of Systems (ANSYS), and Computational Fluid Dynamics (CFD) are a few examples of 3D modeling software that can be used to visualize DTs (Goraj et al., 2019), (Botkina et al., 2018). Submerged simulation is in the third level, which refers to Virtual Reality (VR) (Thomas et al., 2006) and Augmented Reality (AR) (Tadeja et al. (a), 2020). VR is a technique that generates an interactive digital or virtual environment that recognizes the interaction between both the virtual and actual fact, whereas AR is a mixture of the real-world environment and the simulation world that provides virtual assistance to a real scenario. These techniques play a significant role in several specific tasks, such as VR for UAV activities in open terrain (Tadeja et al. (b), 2020). The use of AR for remote aircraft maintenance.

6. Findings & Recommendations

After reviewing various DT-based papers, qualitative and quantitative performance analyses of different models are found. Besides, a few challenges and their possible recommendations are also stated in this section.

6.1 Qualitative Performance Analysis

6.1.1 Planning

(Closed loop, 2022) validated the simulated cargo load plan by comparing it to the load plan generated by the airline's system. Here for performance comparison, the number of available unit load devices, score, number of selected unit load devices, destination, actual loading operation time, and optimization simulation model run time is used. The simulated results show an improvement, particularly in terms of minimizing center of gravity deviations and, as a result, lowering fuel consumption.

6.1.2 Designing

According to (Chang et al., 2021), using the proposed KDT-SD structure, it is probable to use DT to minimize the complexity and reveal design flaws in AAL

design. Furthermore, the knowledge library provides DT with the ability to perform fast modeling and intelligent evaluation, which improves design quality and efficiency.

6.1.3 Modeling

For validation of the model, (Liu et al., 2020) tested their DT model on an air rudder. The DT method recognizes the state of the machine tool, the instrument, the process variables, and other details, and then performs real-time monitoring. According to the authors, their mimic model can include 3D- As-Designed and 3D- In-Process models as well as product specification, structural features, modeling mode, planned process, and background relevant data. But the model-based definition or we can say the physical system cannot provide information on the physical characteristics, simulation mode, planned process, and context information. So, the DT-based mimic model provides more information than the model-based physical system. Authors should work on the speed and accuracy of the model.

(Jinsong et al., 2019) used five performance metrics to evaluate their aerospace-based DT model: execution time, quality inspection batch, total running time, downtimes, and logistic accuracy rate. Here, for comparison, the authors picked the manual+NC machines and the manufacturing using the DT model. Using the manual+NC machines, the execution time, quality inspection batch, running time, downtime, and logistic accuracy rates are 36min, 128 pieces, 43 mins, 1, and 95.5%, respectively. Whereas Using the DT model, the execution time, quality inspection batch, running time, downtime, and logistic accuracy rates are 19min, 4 pieces, 23 mins, 0, and 99.2%, respectively. So, it can be stated that the DT model is more efficient than the manual+NC machine.

6.1.4 Manufacturing

For the result analysis, (Zambal et al., 2018) used a modular variant of the LScan (depth data) and FScan (fiber orientation measurement) sensors to cover the full width of the carbon material. Using those sensor data or the proposed manufacturing dataset, the effect of the defects was calculated in near real-time with the help of an analytical model. In the case of severe defects, a finite-element model was required, which generates results in about 5 minutes. This performance was adequate for use in real-world manufacturing.

6.2 Capturing Physical Properties with High Precision

DT is a complicated procedure that requires the collection of a variety of physical parameters. By accurately capturing their material characteristics, simulating their behavioral patterns, and adjusting their scale, DT technology should be capable of

simulating both small items (such as a vehicle, aircraft, building, or a living person) and complicated relationships (such as a shopping process, irrigation facilities, a store, a metropolitan area, and so on). Even noise is also a factor that affects the quality of DT. Modern Multi-Loop Feedback filters can help to reduce noise from data. Noise removal can also be accomplished using Real-time noise reduction using deep learning. Instead of using a single sensor, we can use multiple sensors for collecting data and the weighted average of those data can be used to increase the accuracy of physical properties.

6.3 Real-Time Automatic Updating

DT is a simulation-based model that runs in real time. It is extremely difficult to collect and process all data in real time. Because of the Internet of Things, this process has become easier in recent years. Nonetheless, it is a really difficult task. The DT model should be self-aware of changes in its size and functionality in real time and adjust or adapt as necessary. Researchers can use a heterogeneous multicore based on Field Programmable Gate Arrays (FPGAs) running on the ReconOS (the operating system for reconfigurable computing) to solve this problem.

6.4 Collaboration on a Project

The DT technology should be able to collaborate on a wide range of projects. Digital twins, for example, can efficiently work within a merged DT. A digital twin can collaborate and link with other DT modules in a variety of settings. By using data from various connected sensors to express the story of a module or asset through its complete life cycle, a DT can create collaboration between different modules. The usage of the shared dataset can also solve this obstacle.

6.5 Interaction with Physical and Digital Objects

Interacting between physical and virtual schemes is highly tough because DT collects data from numerous physical infrastructures and uses that info to operate in a virtual environment. It can take on the look of a human user and communicate with other virtual items or digital twins in an immersive manner. A data interaction mechanism that combines message middleware, relational databases, and memory databases can be configured to support virtual-physical interaction. To send production and manufacturing instructions, responses, and execution results message middleware such as Kafka and MQTT (MQ Telemetry Transport) can be used. Throughout production, a memory database, such as Redis (Remote Dictionary Server), can be used to capture real-time operating data. Information management for simulation models can be done with relational databases such as Oracle, MySQL, SQL Server, and others.

6.6 Detecting and resolving conflicts

The DT model is designed to detect and respond to changes in a real-world setting. To ensure that such interactions are accurate, it must also be able to recognize and correct execution differences. A conflict checker, which indicates possible normative conflicts, can be used to overcome this barrier. Figure 3 shows an overview of all the possible challenges and recommendations.

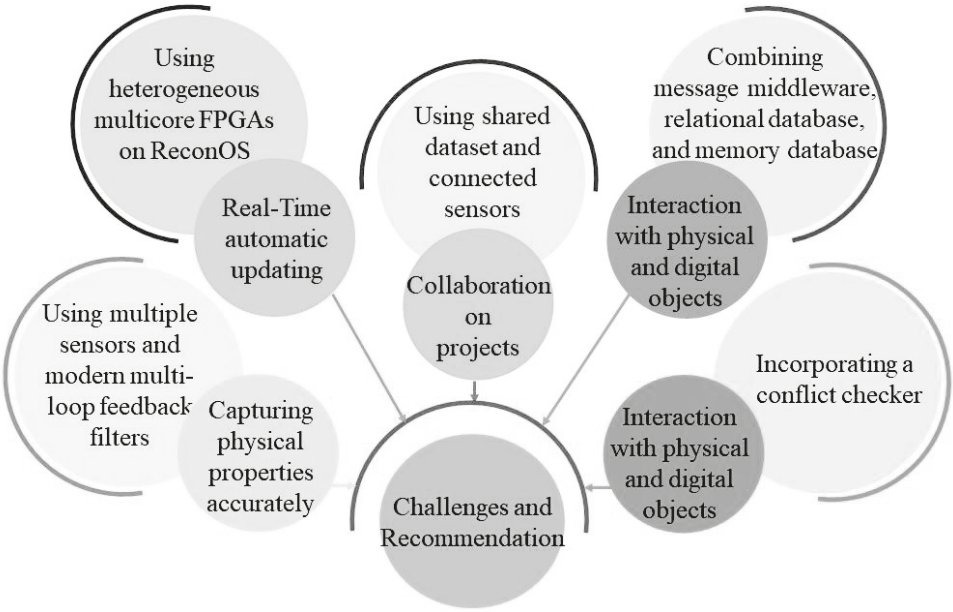


Figure 3: Challenges of DT base models in the Aerospace Industry along with possible recommendations.

7. Conclusion

Today, the DT is an important phenomenon that exists, starting from the design process to maintenance and management. The Digital Twin is also associated with the need to defend the item throughout its life cycle. When the item or object is used, the digital copy, which is fed with simulation studies and design information, should strive to evolve with data obtained through sensors and user inputs from the product's application domains. This comprehensive review is fully focused on the applications of various DT models in the aerospace industry. We have reviewed different useful DT models for designing manufacturing, monitoring, management, and performance optimization in the aerospace industry field. We

have also provided a clear concept of the history and research fields of DT. A few model-based DT models and simulation techniques are also discussed in this paper. One limitation of our work is that we could not describe specific DT models for important parts of aircraft. After reviewing all models, we concluded that DT would help with future versions of aerospace production. So, improvements must be made by reducing the number of copies and errors. Simultaneously, by running numerous tests on the virtual copy, solutions to current mistakes can be found safely. We hope this comprehensive review will help future researchers to perform revolutionary research in this field. In the future, we want to focus on artificial intelligence-based DT models in the aerospace industry.

References

A digital twin approach for rotorcraft flight parameter optimization to enhance component life | *AIAA SciTech Forum*. (n.d.). Retrieved July 3, 2022, from https://arc.aiaa.org/doi/10.2514/6.2021-0899

Aivaliotis, P., Georgoulias, K., Arkouli, Z., & Makris, S. (2019). Methodology for enabling Digital Twin using advanced physics-based modeling in predictive maintenance. *Procedia CIRP*, *81*, 417–422.

Aydemir, H., Zengin, U., &Durak, U. (2020). The Digital Twin Paradigm for Aircraft Review and Outlook. In *AIAA Scitech 2020 Forum*. American Institute of Aeronautics and Astronautics.

Bäßler, M., Haufe, S., Meyer, H., Zimdahl, J., Kamtsiuris, A., Meissner, R., & Raddatz, F. (2020, November 25). *DEVELOPMENT OF A DIGITAL TWIN FOR AVIATION RESEARCH*. https://doi.org/10.25967/530329

Boeing CEO Talks "Digital Twin" Era of Aviation. (2018, September 14). Aviation Today. https://www.aviationtoday.com/2018/09/14/boeing-ceo-talks-digital-twin-era-aviation/

Botkina, D., Hedlind, M., Olsson, B., Henser, J., &Lundholm, T. (2018). Digital Twin of a Cutting Tool. *Procedia CIRP*, *72*, 215–218.

Breaking New Ground: Digital Twin Helps Engineers Design Megawatt-Sized Circuit Breakers | GE News. (n.d.). Retrieved July 3, 2022, from https://www.ge.com/news/reports/breaking-new-ground-digital-twin-helps

Brusa, E. (2021). Digital Twin: Toward the Integration Between System Design

and RAMS Assessment Through the Model-Based Systems Engineering. *IEEE Systems Journal*, *15*(3), 3549–3560.

Cai, H., Zhu, J., & Zhang, W. (2021). Quality Deviation Control for Aircraft Using Digital Twin. *Journal of Computing and Information Science in Engineering*, *21*, 1–17.

Chang, X., Jia, X., Liu, K., & Hu, H. (2021). Knowledge-enabled digital twin for smart designing of aircraft assembly line. *Assembly Automation*, *41*(4), 441–456.

Cimino, C., Negri, E., &Fumagalli, L. (2019). Review of digital twin applications in manufacturing. *Computers in Industry*, *113*, 103130.

Closed-loop digital twin system for air cargo load planning operations. (n.d.). Retrieved July 3, 2022, from https://www.tandfonline.com/doi/full/10.1080/0951192X.2020.1775299

Digital twin-driven optimization of gas exchange system of 2-stroke heavy fuel aircraft engine—ScienceDirect. (n.d.). Retrieved July 3, 2022, from https://www.sciencedirect.com/science/article/abs/pii/S0278612520301357

Digital Twins Grow Up. (2021, April 29). Techonomy. https://techonomy.com/digital-twins-grow-up/

Dröder, K., Bobka, P., Germann, T., Gabriel, F., & Dietrich, F. (2018). A Machine Learning-Enhanced Digital Twin Approach for Human-Robot-Collaboration. *Procedia CIRP*, *76*, 187–192.

Ezhilarasu, C. M., &Jennions, I. K. (2020). A System-Level Failure Propagation Detectability Using ANFIS for an Aircraft Electrical Power System. *Applied Sciences*, *10*, 2854.

Gavin, J. G. (1982). The Aerospace Industry Viewpoint. In *Aerospace Technology and Commercial Nuclear Power* (pp. 39–44). American Institute of Aeronautics and Astronautics, Inc.

Gelernter, D. (1991). *Mirror Worlds: Or: The Day Software Puts the Universe in a Shoebox...How It Will Happen and What It Will Mean*. Oxford University Press.

Glaessgen, E. H., &Stargel, D. S. (2012, April 16). *The Digital Twin Paradigm for Future NASA and U.S. Air Force Vehicles*. 53rd

AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference - Special Session on the Digital Twin, Honolulu, HI.

Goraj, R. (2020). Digital twin of the rotor-shaft of a lightweight electric motor during aerobatics loads. *Aircraft Engineering and Aerospace Technology*, *92*(9), 1319–1326.

Grieves, M. W. (2019). Virtually Intelligent Product Systems: Digital and Physical Twins. In *Complex Systems Engineering: Theory and Practice: Vol. Volume 256* (pp. 175–200). American Institute of Aeronautics and Astronautics, Inc.

Hänel, A., Schnellhardt, T., Wenkler, E., Nestler, A., Brosius, A., Corinth, C., Fay, A., &Ihlenfeldt, S. (2020). The development of a digital twin for machining processes for the application in aerospace industry. *Procedia CIRP*, *93*, 1399–1404.

Ibrion, M., Paltrinieri, N., &Nejad, A. R. (2019). On Risk of Digital Twin Implementation in Marine Industry: Learning from Aviation Industry. *Journal of Physics: Conference Series*, *1357*(1), 012009.

Jeon, H. Y., Justin, C., &Mavris, D. N. (n.d.). Improving Prediction Capability of Quadcopter Through Digital Twin. In *AIAA Scitech 2019 Forum*. American Institute of Aeronautics and Astronautics.

Jinsong, B., Guo, D., Li, J., & Zhang, J. (2018). The modelling and operations for the digital twin in the context of manufacturing. *Enterprise Information Systems*, *13*, 1–23.

Keivanpour, S. (2022). A Conceptual Framework for Estimating the Remaining Operational Lifetime of the Recovered Components from End of Life Aircraft Using Fuzzy Simulation and Digital Twin. In C. Kahraman& S. Aydın (Eds.), *Intelligent and Fuzzy Techniques in Aviation 4.0: Theory and Applications* (pp. 301–316). Springer International Publishing.

Kraft, E. M. (2018). Developing a Digital Thread / Digital Twin Aerodynamic Performance Authoritative Truth Source. In *2018 Aviation Technology, Integration, and Operations Conference*. American Institute of Aeronautics and Astronautics.

L. Li, S. Aslam, A. Wileman, and S. Perinpanayagam, “Digital Twin in Aerospace Industry: A Gentle Introduction,” IEEE Access, vol. 10, pp. 9543–9562, 2022.

Leser, P. E., Warner, J. E., Leser, W. P., Bomarito, G. F., Newman, J. A., & Hochhalter, J. D. (2020). A digital twin feasibility study (Part II): Non-deterministic predictions of fatigue life using in-situ diagnostics and prognostics. *Engineering Fracture Mechanics*, *229*, 106903.

Li, L., Aslam, S., Wileman, A., &Perinpanayagam, S. (2022). Digital Twin in Aerospace Industry: A Gentle Introduction. *IEEE Access*, *10*, 9543–9562.

Li, C., Mahadevan, S., Ling, Y., Wang, L., &Choze, S. (2017, January 9). *A dynamic Bayesian network approach for digital twin*.

Liang, B., Liu, W., Liu, K., Zhou, M., Zhang, Y., & Jia, Z. (2020). A Displacement Field Perception Method for Component Digital Twin in Aircraft Assembly. *Sensors*, *20*(18), 5161.

Liu, S., Bao, J., Lu, Y., Li, J., Lu, S., & Sun, X. (2021). Digital twin modeling method based on biomimicry for machining aerospace components. *Journal of Manufacturing Systems*, *58*, 180–195.

Liu, Z., Meyendorf, N., &Mrad, N. (2018). The role of data fusion in predictive maintenance using digital twin. *AIP Conference Proceedings*, *1949*(1), 020023.

M. Singh, E. Fuenmayor, E. Hinchy, Y. Qiao, N. Murray, and D. Devine, "Digital Twin: Origin to Future," ASI, vol. 4, no. 2, p. 36, May 2021.

Mandolla, C., Petruzzelli, A. M., Percoco, G., &Urbinati, A. (2019). Building a digital twin for additive manufacturing through the exploitation of blockchain: A case analysis of the aircraft industry. *Computers in Industry*, *109*, 134–152.

Model-Based Design of Complex Aeronautical Systems Through Digital Twin and Thread Concepts | IEEE Journals & Magazine | IEEE Xplore. (n.d.). Retrieved July 3, 2022, from https://ieeexplore.ieee.org/document/8835120

NASA Technology Roadmaps—TA 12 Materials, Structures, Mechanical Systems, and Manufacturing. (2015). *Mechanical Systems*, 138.

Phanden, R. K., Sharma, P., & Dubey, A. (2021). A review on simulation in digital twin for aerospace, manufacturing and robotics. *Materials Today: Proceedings*, *38*, 174–178.

Reifsnider, K., & Majumdar, P. (2013, April 8). *Multiphysics Stimulated*

Simulation Digital Twin Methods for Fleet Management. 54th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference.

Rocca, R., Rosa, P., Sassanelli, C., Fumagalli, L., & Terzi, S. (2020). Integrating Virtual Reality and Digital Twin in Circular Economy Practices: A Laboratory Application Case. *Sustainability*, *12*(6), 2286.

Saifutdinov, F., Jackson, I., Tolujevs, J., &Zmanovska, T. (2020). Digital Twin as a Decision Support Tool for Airport Traffic Control. *2020 61st International Scientific Conference on Information Technology and Management Science of Riga Technical University (ITMS)*, 1–5.

Seshadri, B., & Krishnamurthy, T. (2017, January 9). *Structural Health Management of Damaged Aircraft Structures Using Digital Twin Concept*. https://doi.org/10.2514/6.2017-1675

Singh, M., Fuenmayor, E., Hinchy, E. P., Qiao, Y., Murray, N., & Devine, D. (2021). Digital Twin: Origin to Future. *Applied System Innovation*, *4*(2), 36.

Singh, S., Shehab, E., Higgins, N., Fowler, K., Erkoyuncu, J. A., & Gadd, P. (2021). Towards Information Management Framework for Digital Twin in Aircraft Manufacturing. *Procedia CIRP*, *96*, 163–168.

Tadeja, S. K., Lu, Y., Seshadri, P., &Kristensson, P. O. (2020). Digital Twin Assessments in Virtual Reality: An Explorational Study with Aeroengines. *2020 IEEE Aerospace Conference*, 1–13.

Tadeja, S. K., Seshadri, P., &Kristensson, P. O. (2020). AeroVR: An immersive visualisation system for aerospace design and digital twinning in virtual reality. *The Aeronautical Journal*, *124*(1280), 1615–1635.

Tang, W., Xu, G., Zhang, S., Jin, S., & Wang, R. (2021). Digital Twin-Driven Mating Performance Analysis for Precision Spool Valve. *Machines*, *9*(8), 157.

Thomas, J. J., & Cook, K. A. (2006). A visual analytics agenda. *IEEE Computer Graphics and Applications*, *26*(1), 10–13.

Towards an extended model-based definition for the digital twin: Computer-Aided Design and Applications: Vol 15, No 6. (n.d.). Retrieved July 3, 2022, from https://www.tandfonline.com/doi/abs/10.1080/16864360.2018.1462569

Tuegel, E. (2012). The Airframe Digital Twin: Some Challenges to Realization. In *53rd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference*. American Institute of Aeronautics and Astronautics.

Wang, H.-K., Haynes, R., Huang, H.-Z., Dong, L., &Atluri, S. N. (2015). The Use of High-Performance Fatigue Mechanics and the Extended Kalman/Particle Filters, for Diagnostics and Prognostics of Aircraft Structures. *CMES - Computer Modeling in Engineering and Sciences*, *105*, 1–24.

West, T. D., & Blackburn, M. (2017). Is Digital Thread/Digital Twin Affordable? A Systemic Assessment of the Cost of DoD's Latest Manhattan Project. *Procedia Computer Science*, *114*, 47–56.

Wu, J., Yang, Y., Cheng, X., Zuo, H., & Cheng, Z. (2020). The Development of Digital Twin Technology Review. *2020 Chinese Automation Congress (CAC)*, 4901–4906.

Zaccaria, V., Stenfelt, M., Aslanidou, I., &Kyprianidis, K. (2018, June 11). *Fleet Monitoring and Diagnostics Framework Based on Digital Twin of Aero-Engines*.

Zakrajsek, A., & Mall, S. (2017, January 9). *The Development and use of a Digital Twin Model for Tire Touchdown Health Monitoring*.

Zambal, S., Eitzinger, C., Clarke, M., Klintworth, J., &Mechin, P.-Y. (2018). A digital twin for composite parts manufacturing: Effects of defects analysis based on manufacturing data. *2018 IEEE 16th International Conference on Industrial Informatics (INDIN)*, 803–808.

Zhidchenko, V., Malysheva, I., Handroos, H., &Kovartsev, A. N. (2018). Faster than real-time simulation of mobile crane dynamics using digital twin concept. *Journal of Physics: Conference Series*, *1096*, 012071.

Left Blank Intentionally

Phthalate Esters (PAEs): Emerging Organic Pollutant in Aquatic Ecosystems

Raihan Sorker[1*], Shamsunnahar Khanam[1]

Abstract

One of the most produced organic chemicals is known as Phthalate esters (PAEs) which are also excessively used with huge applications in industrial procedures as well as in packaging, medical centers, cosmetics, painting, agriculture, and consumer product uses. These organic pollutants are ubiquitous in the environment, mostly due to uncontrolled urbanization and are a cause of adverse impacts on humans and other living organisms. They even possess carcinogenicity characteristics and endocrine disrupting properties in the aquatic environment through their metabolites. Several studies were found to report the disrupting impacts these compounds make with their prevalence, toxicity, and exposure paths in the aquatic system, including humans. Waters are specifically vulnerable to PAEs because of the various sources of input through land runoff, agriculture, urban households, leaching of wastes etc., thus making it an emerging pollutant in the water. However, modern methodologies and instrumentations have been successful to measure even small amounts of PAEs in lakes, rivers, oceans, and other samples of aquatic systems. This study aims to *provide a thorough study of the distribution and characteristics of phthalate esters and their effects on aquatic ecosystems including aquatic organisms and humans all over the world. These data will be beneficial for understanding the overall distribution of PAEs in the aquatic environment and reducing their ecological footprint.*

Keywords: PAEs, Aquatic ecosystem, Endocrine disruptor, Drinking water, Surface water, Wastewater.

1. Introduction

Phthalate esters (PAEs) are organic micro pollutants which have great public concern due to their serious ecological and human health effects. Phthalate Esters are used in industries that have a common chemical structure of dialkyl or alkyl/aryl esters of 1,2-benzenedicarboxylic acid (Annamalai and Vasudevan, 2020; Giuliani et al., 2020). Phthalic anhydride and suitable alcohol produce phthalic acid's easters which are known as PAEs (Howdeshell et al., 2008; Vats et

[1] Department of Environmental Science, Bangladesh University of Professionals, Dhaka 1216, Bangladesh.

* Correspondence: raihansorker17@gmail.com

al., 2020; Giuliani et al., 2020). Phthalic anhydride and suitable alcohol produce phthalic acid's easters which are known as PAEs (Howdeshell et al., 2008; Vats et al., 2013). The polar carboxyl group only provides some characteristics of PAEs if the small alkyl groups, generally presented by R and R' such as methyl or ethyl are present.

Some of the PAEs that are used on a daily basis in usable products include (DMP), benzylbutyl phthalate (BBP), diisobutyl phthalate (DIP), di-n-butyl phthalate (DBP), diethyl phthalate (DEP), di-n-octyl phthalate (DOP), dicyclohexyl phthalate (DCHP), bis (2-ethylhexyl) phthalate (DEHP), diisodecyl phthalate (DiDP), di-n-hexyl phthalate (DHP), and diisononyl phthalate (DiNP) (Wang et al., 2019; Dutta et al., 2020; Zhang et al., 2021).

PAEs also show a lower melting point and greater boiling points which contributes to their heat-transfer abilities (Giuliani et al., 2020). PAEs can be linear and branched types, both of which are used in plastic production. Less volatilities are the reason for linear structures to have extraordinary flexibilities, and volatility being a concern, lower alkyl chain containing PAEs are not often used to produce plastics. Many daily activity items can contain PAEs including paint, vinyl flooring, printing ink, adhesive, packaging materials, capsules for medicine, cosmetics, packaging etc (Giuliani et al., 2020; Zhang et al., 2021).

PAEs are increasing their commercial needs, as a result, they are being used at a great amount since the 1930s (Wang et al., 2018; Heo et al., 2020) with an estimated global production of 6-8 million tons annually (Seyoum and Pradhan, 2019). The primary use of the PAE compounds can be found in enteric coatings, pills, stabilizer film, lubricants, emulsifying agents, suspending agents, dispersants, viscosity controlling agents and many more; mainly they are utilized as plasticizer (Ji et al., 2014; Wang et al., 2015; Wang et al., 2018; Heo et al., 2020; USEPA, 2021). PAEs are used by adding these substances to different plastic materials. Some of the plastic materials that contain at best 60% of PAEs are polyvinyl chloride (PVC), polyethylene terephthalate, polyethylene, polyvinyl acetate, etc. They are beneficial to increase the plasticity and extensibility of plastic polymers (Giuliani et al., 2020). Some other uses of PAEs are in building equipment, agricultural products, medical instruments, toys for children, personal and health care products, waxes, textiles, glues and adhesives, etc. (Ling et al., 2007; Gao and Wen, 2016). Modern electronic devices like blood transfer and catheters were also reported to contain PAE's in some research (Chakraborty et al., 2019, Malarvannan et al., 2019). As the usage of PAE is increasing in different segments of the consumer cycle, it has gained the spotlight in research and studies

have confirmed the presence of PAEs in different environments (Salaudeen et al., 2018; Ai et al., 2021; Li et al., 2021).

In this paper, the main focus is on the presence and effects of PAEs in aquatic environments. These compounds are known for their bioavailability, and they possess degradable nature, as a result, there are great possibilities of PAE accumulation in higher organisms as well as aquatic life (Arambourou et al., 2019; He et al., 2020; Zhang et al., 2021). Several secondary sources were used to conduct this review. Most recent reports and research covering the prevalence of PAEs and their adverse impacts as well as historical studies were analyzed to compile this review. Worldwide impacts of PAEs on the aquatic ecosystems were studied. Majority of the scientific journal articles and reports were taken from time between 2006 and 2022, but the recent (within 5 years) articles were prioritized. More than 2000 journal articles and reports were found using some keywords such as phthalate esters, phthalate esters in water, phthalate esters toxicity, treatment of phthalate esters, phthalate esters in aquatic ecosystems, phthalate esters bioaccumulation, phthalate esters in sediments, phthalate esters biodegradation, etc. Although not all the articles were directly relevant to the study, thus, around 79 articles were thoroughly read and used in the study. This study has highlighted the current knowledge of PAEs occurrence, ubiquity to identify impacts and toxicity of PAEs in aquatic ecosystems and human health. The aim of this research is to provide better knowledge and understanding toward the effects and distributions of PAEs as emerging organic pollutant in the aquatic environment.

2. Methodology

2.1 Data collection strategies

To reflect the overall condition of PAEs in the aquatic ecosystem a systematic literature review was conducted using the electronic search of Google scholar, ScienceDirect, and PubMed.

In addition, international monitoring organizations such as United Nations Environmental Program (UNEP), United States Environmental Protection Agency (USEPA), and World Health Organization (WHO) reported data was analyzed to know PAEs prevalence, impacts, reduction, and regulation strategies of different countries. From the large number of literatures, this review focused on studies from 2006 to 2022 but priorities in recent five years on aquatic ecosystem in different countries around the world. To exhibit the ubiquity of PAEs in aquatic ecosystem

18 countries PAEs level has studied and most of them was from Asia where 16 literatures was about China and 9 was from France which was shown in the map (Figure 1). Furthermore, the effects and toxicity of PAEs in environment and human health studied from several literatures were reported.

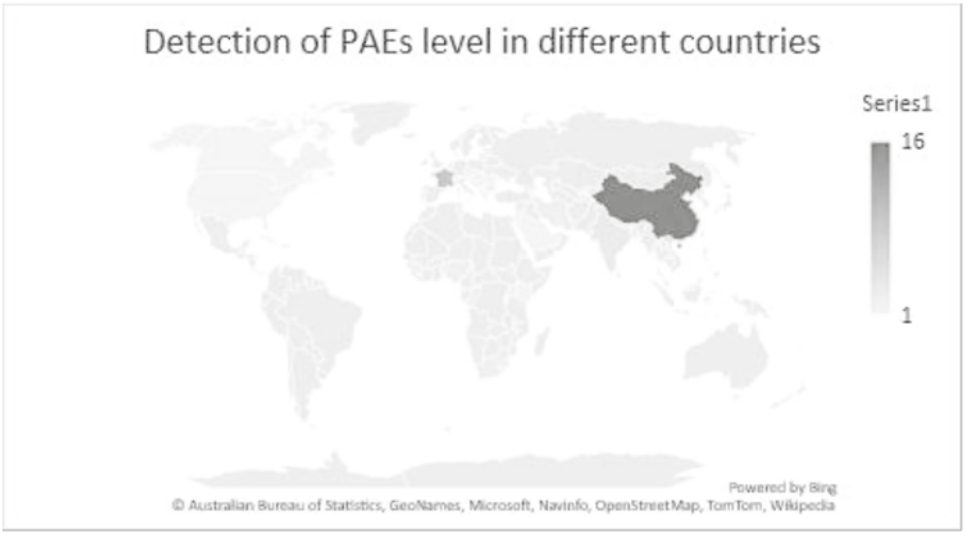
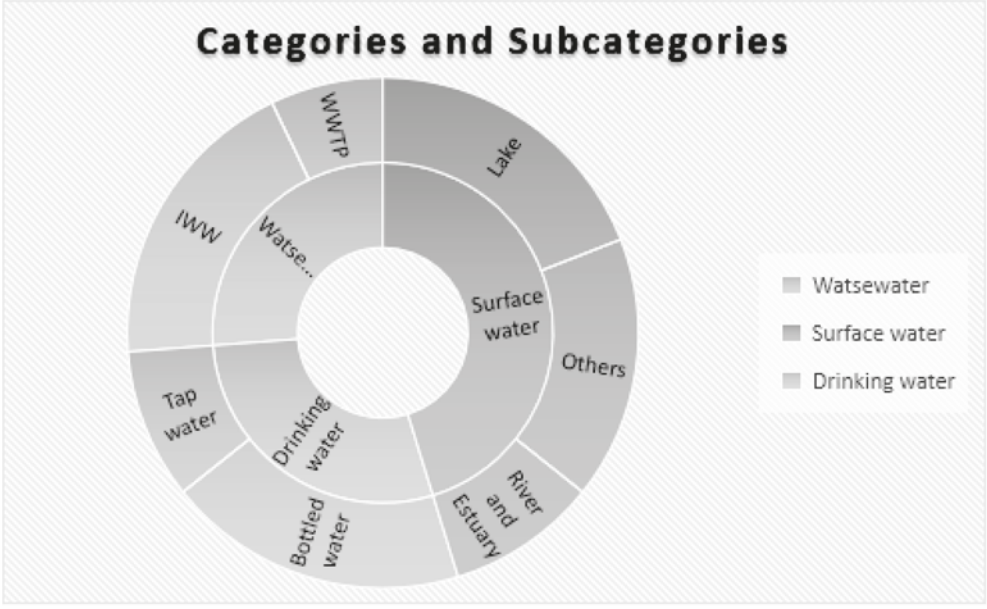


Figure 1: Geographic Locations of Detected Paes Level from Selected Literature



IWW=Industrial Wastewater; WWTP= Wastewater Treatment Plant

Figure 2: Category and Subcategory of Publications

The literature was screened with three categories which divided into seven subcategories (Figure 2). The selection criteria of the literature focused on those publications which was followed by authentic analytical methods and determined the concentration of PAEs individually. Literature on the concentration of PAEs was found by searching keywords as "Phthalates in Wastewater", "Phthalates in Surface, river and lake water", and "Phthalates in drinking water". Furthermore, impacts and toxicity of PAEs was studied by searching "Impacts and toxicity of PAEs in aquatic Ecosystem", "Impact and toxicity of PAEs in human health". The data collection strategies are depicted in Figure 3.

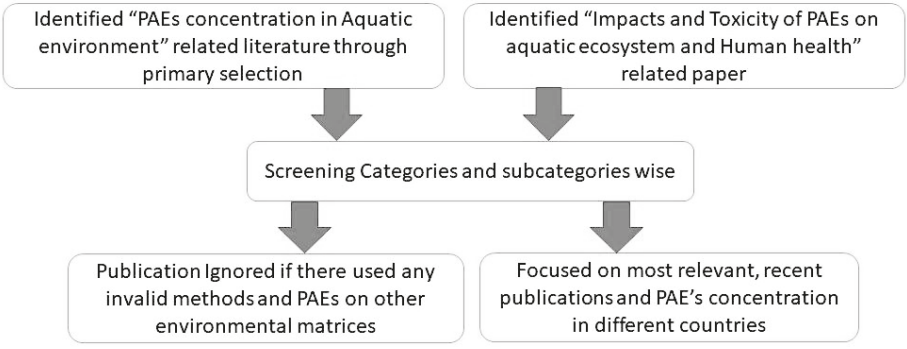


Figure 3: Data Collection Strategies on PAEs

3. Results and Discussion

3.1 Phthalate Esters

Phthalates are synthetic organic chemicals which consists of chemical structure with di alkyl or alkyl/ aryl esters of phthalic acid (Figure 4). Phthalates have low melting point (<-25°C) and it's an odorless chemical which make them a suitable compound as plasticizer (Stales et al., 1997). Therefore, it's not surprising that phthalate esters are ubiquitous compound in aquatic environmental matrices due to their widespread uses in industrial and domestic products.

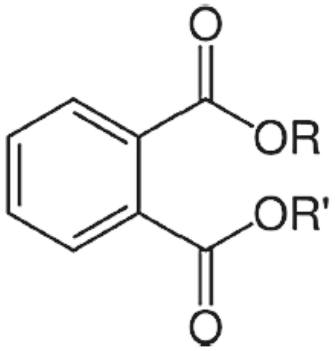


Figure 4: Phthalate esters general structure

This endocrine disruptor released into the aquatic environment via waste disposal, production, use and can easily leached out into water (Salaudeen et al., 2018; Henkel et al., 2019). As plastics is everywhere, plastic aged and degrades which increases the release of PAEs in the environment. Chinese Environmental Agency and United States Environmental Protection Agency selected six targeted PAEs as priority pollutant which included di-(2-ethylhexyl) phthalate (DEHP), dimethyl phthalate (DMP), diethyl phthalate (DEP), dibutyl phthalate (DBP), di-n-octyl phthalate (DnOP), butyl benzyl phthalate (BBP) (Wang et al., 2021). Table-1 depicted the potential application, properties, and toxicity parameters of six targeted phthalate.

Table-1: Physicochemical properties and potential application of six targeted phthalate (Wang et al., 2021, Baloyi et al., 2021)

PAEs	CAS	Molecular mass (g/mol)	RfD (mg/ kg/day)	Carcinogenic grade	Potential application
DEHP	117-81-7	390.56	0.02	B2	Medical devices, Tubing, blood storage bag, plastic toy, wall tiles, baby pants, dolls, food packaging, wire, cables, furniture materials.
DMP	131−11−3	194.18	10	D	Insecticides, adhesive, shampoo, air freshener, hair styling products.
DEP	84−66−2	222.24	0.8	D	Cosmetics, perfumes, deodorant, nail polish, pill coating,

					pharmaceuticals, Aftershave, makeup, lotion.
DBP	84−74−2	278.34	0.1	D	Nitrocellulose lacquers, explosives, nail polish, textile lubricating agent, printing ink.
BBP	85−68−7	312.36	0.2	C	Adhesive, vinyl flooring, carpet tiles, food packaging, and artificial leather.
DnOP	117−84−0	390.56	0.02	-	Mostly used in Children's toy. Other uses as dye carrier, plastisol, colourants.

RfD= reference dose for oral exposure; CAS= chemical abstracts service

Like other organic compounds, PAEs also lipophilic and extracted with lipid soluble chemicals. To detect PAEs in the aquatic environment EPA has provided standard methods (Table 2). Liquid Liquid Extraction (LLE) and Solid Phase Extraction (SPE) is the part of this method as sample preparation process. In SPE process, analytes are isolated from mixture in response to its physicochemical properties (Falenas, 2019). On Contrary, LLE is a basic technique uses two different immiscible liquids such as water and organic solvent according to the preference. Moreover, SPE technique uses in extraction process is increasing because of its easy application, ability to save solvent and time (Net et al., 2015). After extracting the analytes, GC method combined with mass spectroscopy (MS)/ flame ionization detectors (FID)/ Electron capture detector (ECD) used as a main tool for the identification and quantification of PAEs (Yang et al., 2015; Salaudeen et al., 2018; Weizhen et al., 2020)

Table-2: Standard method of detection PAEs level by EPA

Water Matrices	Extraction	Solvent	Quantification	EPA method No
Drinking water	LLE, SPE	Dichloromethane (DCM)	GC-MS	506, 525.2
Municipal and industrial wastewater	LLE	Dichloromethane (DCM)	GC-MS, GC-ECD	606, 625

3.2 Ubiquity of phthalate esters in aquatic ecosystem

In aquatic environment, PAEs concentration found trace µg/L to mg/L varying on their degree of uses in products all over the world. Also, the large variety and abundance of PAEs have been observed. However, the pollution and occurrence level mostly were mild to moderate. Furthermore, Previous studies indicate that most dominant PAEs in the aquasphere are DEHP, DBP, DMP, DEP, and BBP (Baloyi et. 2021; Luo et al., 2021; Zhao et al., 2020; Zheng et al., 2014).

3.2.1 Wastewater

Wastewater is one of the main hotspots of contamination of PAEs as industrial and domestic uses of PAEs in products end up in wastewater and transferred into different water matrices. Several literatures measured the concentration of PAEs in domestic and industrial wastewater (Table 3). Among all studied literatures, it has been observed that DEHP is higher in concentration in both wastewater where the highest amount found in Paris conurbation catchment WWTP (39.4-160 µg/L) and in Textile (38.8-248 µg/L). In Ikeja WWTP, Nigeria found greater level of PAEs including DEHP (15.7-43.5 µg/L), DMP (10.8- 24.9 µg/L), DEP (9.86- 16.6 µg/L), DBP (17.7- 25.3 µg/L), and BBP (0.52- 10.8 µg/L). PAEs were found in textile, pharmaceuticals, Aerospace, and cosmetic industries. PAEs in wastewater were found in several countries like India, France, China, Vietnam, Saudi Arab, South Africa, and Vietnam (Table 3). However, numerous countries didn't study the PAEs level in their wastewater yet.

Table-3: Concentration of PAEs in Wastewater in different countries

Matrices	Concentration in µg/L						Country	Ref
	DEHP	DMP	DEP	DBP	DnOP	BBP		
Wastewater Treatment Plants (WWTP)								
North Indian WWTP	n.d.-132	-	n.d.-20	0.8-90	-	n.d.-9	India	Gani et al., 2016
Paris conurbation catchment WWTP	39.4-160	-	5.23-17.7	0.60-3.31	-	0.46-3.91	France	Bergé et al., 2014
Harbin WWTP	1.7-25.4	n.d.-1.52	n.d.-1.37	3.47-4.13	1.11-14.15	n.d.-17.03	China	Gao et al., 2014
Fontenay-lesBriis WWTP	2.0 ± 1.2	0.03 ± 0.03	0.04 ± 0.05	0.14 ± 0.10	0.01 ± 0.01	0.16 ± 0.15	France	Tran et al., 2015

Riyadh WWTP	0.11-2.85	0-0.81	0-0.63	0.045-3.02	0-0.58	0-1.04	Saudi Arab	Al-Saleh et al., 2017
Ikeja WWTP	15.7-43.5	10.8-24.9	9.86-16.6	17.7-25.3	-	0.52-10.8	Nigeria	Olujimi et al., 2017
Adelaide WWTP	3.44−48.16	1.35-12.07	2.53-24.42	nd−45 1.48	nd−21.75	2.38−80.70	South Africa	Salaudeen et al., 2018
Hanoi WWTP	2.84-42.4	0.22-1.87	1.26–7.04	1.58–9.45	0.53-2.73	0.380–8.97	Vietnam	Le et al., 2021
Industrial Wastewater								
Textile	38.8-248	-	2.87-76.3	2.26-14.9	-	0.44-6.08	France	Bergé et al., 2014
Pharmaceutical	0.60-386	-	0.47-137	0.48-6.96	-	0.13-2.90	France	Bergé et al., 2014
Aerospace industries	0.80-33.3	-	0.22-4.85	0.44-1.19	-	0.03-2.21	France	Bergé et al., 2014
Cosmetic industries	6.04-85.2	-	112-120	0.37-1.19	-	0.04-0.04	France	Bergé et al., 2014
Domestic and industrial	4.1-13	0.26-0.81	n.d.-2.7	<0.1-0.47	n.d.-< 0.1	n.d.-0.11	Austria	Clara et al., 2010

3.2.2 Surface water

The level of PAEs was detected in surface water among different countries (Table-4). Unlike the concentration in wastewater the level of PAEs is lower in surface water. Adeniyi et al. (2011) detected greater level of PAEs (255-2705 µg/L) in Ogun River, Nigeria among all other studies. On contrary, Chepchirchir et al. (2017) found mild level of PAEs (0.001-0.1 µg/L) in Sosiani River, Kenya. Studies showed that frequently the standard level of annual DEHP (1.3 µg/L) in surface water exceeded (EU directive 2008/105/EC) (UNEP, 2020). PAEs detection and studies rate is higher in Asia due to not having restrictions in using PAEs unlike in developed countries such as USA and Europe (UNEP, 2020; Das et al., 2021; Li et al., 2021). Surface water is more contaminated with PAEs which are adjacent to industries and other notable anthropogenic activities such as manufacturing, Agriculture, solid waste disposal and wastewater treatment plants (Zhu et al., 2019).

Table-4: Concentration of PAEs in Surface water in various countries

Matrices	Concentration in μg/L						Country	Reference
	DEHP	DMP	DEP	DBP	DOP	BBP		
River water								
Hanjiang River	0.36-0.42	0.02-0.04	0.13-0.25	0.001-0.59	0.01-0.05	0.003-0.02	China	Dong et al., 2022
Yellow River	0.43-0.83	0.38-0.64	0.31-0.45	0.78-0.80	0.002 - 0.004	0	China	Zhao et al., 2020
Rhone River, France	0.4068	0.0057	0.03 05	0.0405	n.d.	-	France	Paluselli et al., 2018
Jiulong River estuary	0.57-3.66	0.03-0.12	0.03-0.05	0.37-0.67	-	-	China	Li et al., 2017
Sosiani River	0.001-0.010	-	-	n.d.-0.140	-	0.002-0.008	Kenya	Chepchirchir et al., 2017
Huai River	0.08-1.52	0.02-0.19	0.02-0.92	2.17-21.98	-	0.02-2.16	China	Shi et al., 2016
Songhua River	2.26-11.55	0.98-4.12	1.33-6.67	1.69-11.81	0.69-6.14	nd-4.39	China	Gao and Wen, 2016
Kaveri River	0.51	0.02	0.24	0.03	0.25	0.04	India	Selvaraj et al., 2015
Pearl River	0.57	0.57	0.22	0.41	0.03	0.19	China	Li et al., 2014
Somme River	5.16-20.80	0.02-0.25	0.26-6.98	0.22-3.86	-	-	France	Net et al., 2014
Jukskei River	0.49-5.58	0.04-0.56	0.08-0.39	-	0.79-3.65	-	South Africa	Sibali et al., 2013
Selangor River	0.39	0.01	0.04	0.21	0.009	0.01	Malaysia	Santhi and Mustafa, 2013
Ogun river	255-480	n.d.	1480 - 1755	2025-2705	-	-	Nigeria	Adeniyi et al., 2011
Seine River Estuary	0.16-0.31	0.03-0.18	0.07-0.18	0.07-0.32	-	-	France	Dargnat et al., 2009
Lake and other surface water								
Hanoi Lake Water	1-48.7	0.11-2.95	0.64-14	0.78-34	n.d. – 7.31	0.18-21.1	Vietnam	Le et al., 2021
Taihu Lake	0.14-0.83	0.05-0.1	0.02-0.09	0.03-1.31	n.d.-0.10	0.04-0.16	China	Luo et al., 2021

Small Xingkai Lake	0.12-3.25	0.001-0.011	0.002-0.007	0.10-0.53	n.d.-0.004	0.001-0.005	China	Yang et al., 2021
Large Xingkai Lake	0.22–3.46	0.003-0.026	0.003-0.018	0.11–1.52	n.d.-0.007	0–0.002	China	Yang et al., 2021
Poyang Lake	0.023-0.896	n.d.-0.253	n.d.-0.127	0.121-1.297	nd-0.018	-	China	Ai et al., 2021
Lake Victoria	0.21–23	0.07–0.4	0.04–1.1	0.35–16	-	-	Uganda	Nantaba et al., 2021
U-Tapao canal	1.28-5.28	-	-	nd-3.36	-	-	Thailand	Kingsley and Witthayawirasak, 2020
Asan Lake	n.d.-1.34	n.d.-0.18	n.d.-0.05	n.d.-0.34	n.d.-0.02	n.d.	Korea	Lee et al., 2019
Lakes Shichaha	0.140-0.519	0.047-0.143	0.006-0.013	0.009-0.157	0.015-0.022	n.d.- 0.512	China	Zheng et al., 2014
Guanting Reservoir	0.043-0.149	0.023-0.084	n.d.-0.006	n.d.-0.594	0.013 0.029	n.d.- 1.246	China	Zheng et al., 2014
Lake Chaohu	n.d.-0.58	0.015-3.67	0.006-0.28	0.070-17.53	n.d.-0.045	n.d.-0.11	China	He et al., 2013
False Creek Harbor	0.17-0.44	0.002-0.005	0.05-0.35	0.05-0.244	0.005-0.04	0.002-0.006	Canada	Mackintosh et al., 2006

3.2.3 Drinking water

It's a concerning issue that several studies found hazardous phthalates in drinking water (Table 5). Yousefi et al. (2019) measured that greater concentration of PAEs in bottled water in Iran and Abdolahnejad et al, (2019) found mild concentration in tap water in Isfahan, Iran. Bottled water contains higher concentration of PAEs than tap water (Table 5). Recent studies on PAEs in bottled water is increasing due to have high exposure especially DEHP is continuously exceeding its RfD value (Baloyi et. al., 2021; Abdolahnejad et al., 2019; Wang and Qian, 2021). DBP and DEHP are the most dominant compounds in drinking water among other phthalate esters (Yousefi et al., 2019). Loraine and Pettigrove (2006) and Liu et al. (2015) investigated PAEs concentration from the source of drinking water and found higher amount of DEHP and DBP.

Table-5: Concentration of PAEs in drinking water of different countries

Matrices	Concentration in μg/L						Location,Country	Ref
	DEHP	DMP	DEP	DBP	DnOP	BBP		
Bottled Water	0.257	0.38	0.198	0.317	0.248	0.633	South Delhi, India	Das et al., 2014
water filtration plants	2.67-5.94	0.098-0.78	0.899-1.49	1.44-8.34	-	0.053-1.19	San Diego, USA	Loraine and Pettigrove, 2006
Tap water	1.01–14.5	n.d. – 0.54	n.d. – 2.57	0.014 –2.56	n.d. – 1.93	0.197 –4.21	Hanoi, Vietnam	Le et al., 2021)
Bottled water	0.23-1.95	n.d. - 0.17	n.d. - 0.99	0.14-3.07	n.d. - 0.94	0.25-4.37	Hanoi, Vietnam	Le et al., 2021
Tap water	0.002-6	-	N.d-0.11	n.d.-0.07		n.d-0.1	Isfahan, Iran	Abdolahnejad et al., 2019
Bottled water	0.24-0.73	-	-	8.98-11.5	n.d-1.60	n.d.-0.69	Chengdu, China	Yin et al., 2019
Tap water	1.84-2.68	-	-	n.d.-2.04	-	-	Sadao, Thailand	Okpara et al., 2022
Bottled water	6.93	2.23	5.92	6.53	-	-	Sari, Iran	Yousefi et al., 2019
Bottled Water	2.18	-	-	3.23	-	-	Cluj-Napoca, Romania	Wang and Qian, 2021)
Waterworks	5.51	0.76	0.19	1.56	0.24	0.35	Qingdao, China	Liu et al., 2015

3.3 Impact on Aquatic Ecosystem

PAEs are currently discovering as a threat to aquatic habitats. The key attributes of effects on ecosystems are continuous inputs and intrinsic toxicity. Anthropogenic activities can contribute huge number of PAEs to the adjacent aquatic ecosystems like lakes, rivers, canals etc. that are situated or crosses agricultural and industrial areas, as well as urban residents (Salaudeen et al., 2018; Ai et al., 2021; Li et al., 2021). Certain anthropogenic activities act as the primary source of PAEs on aquatic ecosystem by releasing phthalate due to industrial activities, landfills, wastewaters, household wastes regardless of some freshwater algae and cyanobacteria being able to produce monoethylhexyl phthalate under natural states (Zhang et al., 2019; Zhang et al., 2021). PAEs generally possess high octanol-water partition (K_{ow}/log $K_{ow}25$) (1.61–9.46) and low vapor pressures (Pa25) (1.84 × 10−6–0.263), as a result, they are less volatile which makes them

flexible to enter different water bodies and ecosystems (Net et al., 2015; Das et al., 2021; Zhang et al., 2021). The higher molecular weight PAE's tend to show higher sorption to organic matter, which is a result of increased hydrophobicity. The escalation of hydrophobicity happens due to rising alkyl chain length which increases the Kow value (Das et al., 2021; Zhang et al., 2021).

Aquatic organisms are exposed to PAEs followed by their entrance into the system. High nutrient-grade organisms may get exposed to PAEs by ingestion, afterwards they are moved up the food chain; the system is affected my PAEs in several manners. The effects are dependent on their capability to show toxicity to aquatic organisms and ecosystems. Some of the most toxic PAEs are BBP, DEHP, and DBP (Staples et al., 1997; Aarab et al., 2006; European Commission, 2008; Arambourou et al., 2019; Zhang et al., 2021); as a result, they pose the most toxic effects to aquatic organisms (Crafford and Avenant-Oldewage, 2010). A study by Staples et al. (1997) showed that the exposure of these compounds as well as DMP, DEP, DIBP, DAP were responsible to create both acute and chronic effects in organisms including fish, invertebrates and algae. The fact is also supported by Zhang et al. (2021) who showed such effects with symptoms of necrosis, cardiac edema, crooked tails, lack of tactile response etc. More harmful effects such as liver, kidney damage, effects on reproduction etc. were observed on adult life forms (Gao et al., 2018).

PAEs containing lower molecular weight are only toxic in concentrations that are lower than their aqueous solubilities. As the alkyl chain length increases, the toxicity also follows up to 4 atoms of carbon (Staples et al., 1997). Unfavorable effects on aquatic life forms were seen from some studies of ecotoxicity that showed broad range of endpoints which were at much decreased level (ng/L to μg/L) than the recent ones (Oehlmann et al., 2009; Zhang et al., 2021).

The LC50, also known as the median lethal concentration, is the determinant of chemical substance toxicity in organisms (Huang et al., 2020; Zhang et al., 2021). Moreover, Zhang et al. (2021) studied a dose of 96h LC50 range on common carp for the PAEs DEP, DBP, DOP, and DEHP and found similarities with marine flounder and Nile tilapia in the DEP and DBP lethal concentrations. Microalgae possess the ability to degrade PAEs, so as a result some benthic diatom such as *Cylindrotheca closteriums* were found to catalyze the deconstruction of DEP and DBP instead of being affected by those (Li et al., 2015) Although some short chained PAEs have a tendency to bioaccumulate, studies found them to be decomposable in nature (He et al., 2020; UNEP, 2020).

Despite of the PAE's ubiquitousness, studies about those including their exposure route to human and behavior of bioaccumulation failed to gather much information. Although their accumulation rate in biota was explored by eco-toxicological research. A study by Net et al. (2015) presented their results on how larger bioaccumulation factor (BAFs) values can cause the accumulation of PAEs in different species at a higher capacity. Variances of the values were also reported in several studies which directs toward a more detailed and improved system of monitoring of PAEs in the aquatic system (Baloyi et. al, 202). Most vulnerable are those communities who benefit from the waters for fishing, personal use, irrigation use etc. and needs more research and continuous monitoring. The fact was supported by He et al. (2020) in his study about a Chinese river-water dependent village when PAEs were found in livestock meat samples. An important relationship between the biota of the area and adjacent river water was also noted by the research team which points to the possibility of river water influencing PAEs uptake and bioaccumulation (He et al., 2020). PAEs containing molecular weights that are higher than usual have showed extraordinary capability of accumulation in sediments (UNEP, 2020).

3.4 Phthalate Toxicity, Exposure and Effects on Human Health

Humans are continuously exposed to endocrine and metabolic disruptor phthalate esters from different compounds in small quantities from thousands of industrial and household products including cosmetics, toys, foods, packaging materials, beverage and drinking water (He et al., 2020; Abtahi et al., 2019; Gong et al., 2018). Humans are exposed from aquatic ecosystem to this chemical through different routes such dermal skin, inhalation, ingestion, and absorption by consuming foods and drinking water (Giuliani et al., 2020; Abtahi et al., 2019; Paluselli et al., 2018). PAEs arises as global concern due to their toxicity, mutagenic and carcinogenic characteristics to humans on exposure (Gao and Wen, 2016; Li et al., 2020; USEPA, 2021). Exposure to PAEs varying different levels by acting as an endocrine disruptor. Endocrine disruptor referred as active hormone agent that act like hormone in the endocrine system and causes physiological dysfunction (Talia et al., 2021; You and Song, 2021; Darbre, 2019). Most concerning issues about this endocrine disruptor posits that low dose of this chemical can interfere with reproduction, growth and other hormonal mediated process (Zoeller et al., 2014; Gore et al., 2015). Though endogenous hormones are typically present in human body, tiny amount of exogenous hormonal active substances can cause severe impacts. Several studies showed some other effects of PAEs on human exposure which listed in (Table 6).

Table-6: Summarized health effects of PAEs from different studies

Exposure measure	Health effects	Ref	Sex	Age
Urine	Atopic dermatitis	Blakeway et al., 2020	All	All
Urine	Endometriosis	Cai et al., 2019	Female	Adult
Urine	Breast Cancer	Fu et al., 2017	Female	Adult
Multiple	Cardiometabolic risk factors, and Obesity	Golestanzadeh et al., 2019	all	Children
Multiple	Semen quality and sperm DNA damage	Høyer et al., 2018	M	Adult
Multiple	Autism	Jeddi et al., 2016	all	children
Urine	Neurodevelopment disorders	Lee et al., 2018	all	children
NR	Early onset puberty	Poursafa et al., 2015	all	children
Urine	Fertility	Vabre et al., 2017	Female	Adult
Blood	Fetal sex hormone changes, Anogenital distance, hypospadias, cryptorchidism; other congenital malformations	Marie et al., 2015	all	Infant
Indoor Air	Asthma	Li et al., 2017	all	children

Human biomonitoring studies shows that the half-life of PAEs on urine and/or blood plasma of human and rodents is less than 24h (Wang et al., 2019; Giuliani et al., 2020). While DEHP binds with blood plasma protein, biliary excretion, and enterohepatic circulation in humans have been indicated that major elimination pathway of PAE is urinary excretion (Frederiksen et al., 2007; Wang et al., 2019). Jiang et al. (2018) investigated that high urinary phthalate metabolites in pregnant women have experienced lower hemoglobin concentrations, increased blood clotting time, and increased likelihood of anemia (Jiang et al., 2018). Also, PAEs have serious adverse effects on development and functioning of male reproductive organs (Xie et al., 2019).

There are several available literatures based on occurrence and levels of PAEs for suggestions to avoid or minimize the exposure to PAEs. Suggestion to minimize the effects of PAEs with limited evidence are- minimal use of personal care products, using alternative of plastics and can for packaging foods, avoiding nutritional supplement, no micro waved foods, frequent handwash and balanced diet (Koniecki et al., 2011; Braun, 2017; Wang et al., 2019). A study by Chen et al. (2015) explored these PAEs exposure minimizing strategies among 30 young girls and post inventory results found that Phthalate metabolites significantly reduced within one-week intervention period. Frequent hand wash ($P = 0.009$) and

avoiding uses of plastic cups ($P = 0.016$) were most significant among all strategies to reduce phthalate metabolites in urine, particularly, DEHP and mono-n-butyl phthalate (MBP) (Chen et al., 2015). It is also recommended to avoid using untreated river and lake water as those could be major source of exposure to phthalate, especially in rural areas where clean water is limited. Another study in China by He et al. (2020) found that using untreated river water is responsible for 79, 83, and 88% of the estimated PAEs daily intake in toddler, children and adults and revealed an increasing trend with age (He et al., 2020). It's also notable that consuming meats, some vegetables (He et al., 2020) and some corns (Rice, and wheat) (Sun et al., 2015; Xu et al., 2020) serve as a source of PAEs exposure. Therefore, consuming organic farm foods which contains no synthetic products would be the best suggestion to avoid PAEs exposure **through ingestion.**

Table-7: Toxicity of PAEs

Phthalates	MRLs/TDI	Toxicity	Citation
DEHP	100 (MRL-Chronic)	Reproduction	ATSDR 2021
	60 (MRL-Intermediate)		
	50 (TDI)		
DBP	500 (MRL- Acute)	Reproduction, Development	ATSDR 2021, European Food Safety Authority 2019
	10 (TDI)		
DEP	7,000 (MRL-Acute)	Reproduction	ATSDR 2021
	6000 (MRL-Intermediate)	Hepatic	
BBP	500 (TDI)	Reproduction	European Food Safety Authority 2019

UNEP (2020) published an assessment report on chemical and wastes issues on posing human health and environmental risk and reviewed that several countries have taken measures to reduce the PAEs exposure. Denmark has successfully reduced the risk of PAEs exposure by imposing tax on using PVC and PAEs and repealed the enforce in 2019 as part of successful reduction of overall PAEs utilization. The Korean Ministry of Food and Drug Safety and though National Food Safety Standard in China prohibited usages of DBP, BBP, and DEHP. China also banned the use of DBP, BBP, DEHP, DOP, DiNP, and DiDP in textile products. Canadian Government banned in using DEHP in cosmetics and BBP, DBP, DEHP, DOP, DiNP, and DiDP restricted in using (below 1 mg/kg). Other countries like Peru and Columbia prohibited using BBP, DBP, DOP, DEHP, DiNP, and DiDP in plastics for children (UNEP 2020). To minimize the widespread effects of PAEs on health and environment, the concerned world leading

international organizations need to monitor and regulate the PAEs level. Several international organizations like Agency for Toxic Substances and Disease Registry (ATSDR) and European Food Safety Authority EFSA regulated PAEs toxicity by setting limit in the form of Minimal Risk Levels (MRLs) and Tolerable Daily Intake (TDI) (Table-7). United States National Academies of Sciences, Engineering and Medicine (2017) obtained that current toxicity method can identify DEHP hazard but cannot determine the level at which human can be affected. In addition, some references doses and safe limits are set by regulating agency based on animal testing which couldn't be safe limit for human (UNEP, 2020). It's high time to discover methods for monitoring studies on human samples. A survey conducted in 2015–2016 by National Health and Nutrition Examination Survey (NHANES) and found that mostly affected community to high concentration PAEs are people living under the poverty line (USCDC, 2018; USEPA, 2018). Another similar trend found in children who lives under low socio-economic income household (Navaranjan et al., 2019; UNEP, 2020). Now the concerning issue that there is not identified limits of daily intake or minimum risk level for most of the PAE compounds. Another worrying and notable fact that these compounds are not regulated in most of the developing countries; hence their impacts could be catastrophic and they are not measuring the level PAEs in their environmental matrices and no epidemiological report has published yet (Baloyi et al., 2021).

4. Conclusion

Presence and distribution of PAEs in environment, specifically in aquatic systems, is one of the most significant issues which needs instant attention. Various point and non-point sources are responsible to contribute the pollutants in the water, thus making it a great concern and several studies supported the fact by showing the exposure of PAEs on aquatic life as well as humans. Studies have focused on the toxicity and its effects of PAEs on aquatic organisms' organs, tissues, and other parts, which have demonstrated a clear need of protection of the water resources and monitoring. Although the various exposure sources have been brought under light by many researchers, studies on integrated and mixed exposures, effects of different sources were limited. This review was done to gather information about the research that focused on the monitoring and characterization of PAEs, while highlighting their adverse effects on humans and other aquatic organisms in aquatic system. However, it is also observed that implementation of strict regulations on PAE supplies in the aquatic systems of developing countries were very limited. The effort of some countries to regulate and monitor the spread of PAEs by various measures were seen to be ineffective. It is the result of burgeoning population of humans that have caused excessive production of PAE as well as its

ubiquitous applications. Thus, more focused research in this area is a matter of great concern.

References:

Abdolahnejad, A., Gheisari, L., Karimi, M., Norastehfar, N., Ebrahimpour, K., Mohammadi, A., Ghanbari, R., Ebrahimi, A., and Jafari, N., 2019. Monitoring and health risk assessment of phthalate esters in household's drinking water of Isfahan, Iran. International Journal of Environmental Science and Technology, 16(11), pp.7409–7416.

Abtahi, M., Dobaradaran, S., Torabbeigi, M., Jorfi, S., Gholamnia, R., Koolivand, and A., 2019. Health Risk of Phthalates in Water Environment: Occurrence in Water Resources, Bottled Water, and Tap Water, and burden of Disease from Exposure through Drinking Water in Tehran, Iran. Environmental Research, 173, pp.469–479. doi:10.1016/j.envres.2019.03.071

Adeniyi, A. A., Okedeyi, O. O., and Yusuf, K. A., 2011. Flame ionization gas chromatographic determination of phthalate esters in water, surface sediments and fish species in the Ogun river catchments, Ketu, Lagos, Nigeria. Environmental Monitoring and Assessment, 172(1–4), pp.561–569. https://doi.org/10.1007/s10661-010-1354-2

Ai, S., Gao, X., Wang, X., Li, J., Fan, B., Zhao, S., and Liu, Z., 2021. Exposure and tiered ecological risk assessment of phthalate esters in the surface water of Poyang Lake, China. Chemosphere, 262, 127864.

Al-Saleh, I., Elkhatib, R., Al-Rajoudi, T., and Al-Qudaihi, G., 2017. Assessing the concentration of phthalate esters (PAEs) and bisphenol A (BPA) and the genotoxic potential of treated wastewater (final effluent) in Saudi Arabia. Science of the Total Environment, 578, pp.440–451.

Annamalai, J., and Vasudevan, N., 2020. Detection of Phthalate Esters in PET Bottled Drinks and lake Water Using esterase/PANI/CNT/CuNP Based Electrochemical Biosensor. Analytica Chimica Acta, 1135, pp.175–186. doi:10.1016/j.aca.2020.09.041

Arambourou, H., Planelló, R., Llorente, L., Fuertes, I., Barata, C., and Delorme, N., 2019. Chironomus Riparius Exposure to Field-Collected Contaminated Sediments: from Subcellular Effect to Whole-Organism Response. Science of the Total Environment 671, pp.874–882. doi:10.1016/j.scitotenv.2019.03.384

ATSDR, 2021. Minimal Risk Levels (MRLs) for Hazardous Substances. Available at: https://wwwn.cdc.gov/TSP/MRLS/mrlsListing.aspx [Accessed 29 April 2022].

Baloyi, N. D., Tekere, M., Maphangwa, K. W., and Masindi, V., 2021. Insights Into the Prevalence and Impacts of Phthalate Esters in Aquatic Ecosystems. Frontiers in Environmental Science, 9:684190.

Bergé, A., Gasperi, J., Rocher, V., Gras, L., Coursimault, A., and Moilleron, R., 2014. Phthalates and alkylphenols in industrial and domestic effluents: Case of Paris conurbation (France). Science of the Total Environment, 488–489, pp.26–35.

Blakeway, H., Van-de-Velde, V., Allen, V. B., Kravvas, G., Palla, L., Page, M. J., Flohr, C., Weller, R. B., Irvine, A. D., McPherson, T., Roberts, A., Williams, H. C., Reynolds, N., Brown, S. J., Paternoster, L., and Langan, S. M., 2020. What is the evidence for interactions between filaggrin null mutations and environmental exposures in the aetiology of atopic dermatitis? A systematic review. British Journal of Dermatology, 183(3), pp.443–451.

Braun, J. M., 2017. Early-life Exposure to EDCs: Role in Childhood Obesity and Neurodevelopment. Nature Reviews Endocrinology, 13 (3), pp.161–173.

Cai, W., Yang, J., Liu, Y., Bi, Y., and Wang, H., 2019. Association between Phthalate Metabolites and Risk of Endometriosis: A Meta-Analysis. International Journal of Environmental Research and Public Health, 16(19), 3678.

Crafford, D., and Avenant-Oldewage, A., 2010. Bioaccumulation of Non-essential Trace Metals in Tissue and Organs of Clarias gariepinus (Sharp Tooth Catfish) from the Vaal River System-Strontium, Aluminium, lead and Nickel. Water SA, 36 (5), pp.621–640.

Chepchirchir, B. S., Paschke, A., and Schüürmann, G., 2017. Passive sampling for spatial and temporal monitoring of organic pollutants in surface water of a rural-urban river in Kenya. Science of the Total Environment, 601–602, pp.453–460.

Chakraborty, P., Sampath, S., Mukhopadhyay, M., Selvaraj, S., Bharat, G. K., and Nizzetto, L., 2019. Baseline Investigation on Plasticizers, Bisphenol A, Polycyclic Aromatic Hydrocarbons and Heavy Metals in the Surface Soil of the Informal Electronic Waste Recycling Workshops and Nearby Open Dumpsites in Indian Metropolitan Cities. Environmental Pollution, 248, pp.1036–1045.

Chen, C.-Y., Chou, Y.-Y., Lin, S.-J., and Lee, C.-C., 2015. Developing an Intervention Strategy to Reduce Phthalate Exposure in Taiwanese Girls. Science of the Total Environment, 517, pp.125–131.

Clara, M., Windhofer, G., Hartl, W., Braun, K., Simon, M., Gans, O., Scheffknecht, C., and Chovanec, A., 2010. Occurrence of phthalates in surface runoff, untreated and treated wastewater and fate during wastewater treatment. Chemosphere, 78(9), pp.1078–1084.

Dargnat, C., Blanchard, M., Chevreuil, M., and Teil, M. J., 2009. Occurrence of phthalate esters in the Seine River estuary (France). Hydrological Processes, 23(8), pp.1192–1201.

Das, M. T., Ghosh, P., and Thakur, I. S., 2014. Intake estimates of phthalate esters for South Delhi population based on exposure media assessment. Environmental Pollution, 189, pp.118–125.

Das, M. T., Kumar, S. S., Ghosh, P., Shah, G., Malyan, S. K., and Bajar, S., 2021. Remediation Strategies for Mitigation of Phthalate Pollution: Challenges andFuture Perspectives. Journal of hazardous materials 409, 124496.

Darbre, P. D., 2019. The History of Endocrine-Disrupting Chemicals. Current Opinion in Endocrine and Metabolic Research, 7, pp.26–33.

Dong, L., Lin, L., Pan, X., Zhang, S., Lv, Z., and Mi, C., 2022. Distribution Dynamics of Phthalate Esters in Surface Water and Sediment of the Middle-Lower Hanjiang River, China. International Journal of Environmental Research and Public Health, 19(5), 2702.

Dutta, S., Haggerty, D. K., Rappolee, D. A., and Ruden, D. M., 2020. Phthalate Exposure and Long-Term Epigenomic Consequences: A Review. Frontiers in Genetics 11, 405.

EPA method 525.2 (Available at: https://www.epa.gov/sites/ production/files/2015-06/documents/epa-525.2.pdf. [Accessed 21 Jan 2021].

EPA method 506 (Available at: https://www.unitedchem. com/wp-content/uploads/2019/08/EPA_Method_506.pdf. [Accessed 29 April 2022]

EPA method 606 (Available at: https://www.epa.gov/sites/ production/files/2015-09/documents/method_606_1984. pdf. [Accessed 29 April 2022]

EPA method 625 (available at: https://19january2017snapshot. epa.gov/sites/production/files/2015-10/documents/ method_625_1984.pdf. [Accessed 29 April 2022]

European Food Safety Authority, 2019. Update of the Risk Assessment of Di-butyl Phthalate (DBP), Butyl-benzyl-phthalate (BBP), Bis(2-ethylhexyl) Phthalate (DEHP), Di-isononyl Phthalate (DINP) and Di-isodecyl Phthalate (DIDP) for Use in Food Contact Materials. Journal Efsa, 17 (12), pp.1–85.

Falenas, F., 2019. Discussion: Automated Filtration or Really Just Solid Phase Extraction. Journal of chromatography A, 1601, pp.388–389.

Frederiksen, H., Skakkebaek, N. E., and Andersson, A.-M. 2007. Metabolism of Phthalates in Humans. Molecular Nutrition and Food Research, 51, pp.899–911.

Fu, Z., Zhao, F., Chen, K., Xu, J., Li, P., Xia, D., and Wu, Y., 2017. Association between urinary phthalate metabolites and risk of breast cancer and uterine leiomyoma. Reproductive Toxicology, 74, pp.134–142.

Gani, K. M., Rajpal, A., and Kazmi, A. A., 2016. Contamination level of four priority phthalates in North Indian wastewater treatment plants and their fate in sequencing batch reactor systems. Environmental Science: Processes and Impacts, 18(3), pp.406–416.

Gao, D., Li, Z., Wen, Z., and Ren, N., 2014. Occurrence and fate of phthalate esters in full-scale domestic wastewater treatment plants and their impact on

receiving waters along the Songhua River in China. Chemosphere, 95, pp.24–32.

Gao, D.-W., and Wen, Z.-D., 2016. Phthalate esters in the environment: A critical review of their occurrence, biodegradation, and removal during wastewater treatment processes. Science of the Total Environment, 541, pp.986–1001.

Gao, D., Li, Z., Wang, H., and Liang, H., 2018. An Overview of Phthalate Acid Ester Pollution in China over the Last Decade: Environmental Occurrence and Human Exposure. Science of the Total Environment, 645, pp.1400–1409.

Golestanzadeh, M., Riahi, R., and Kelishadi, R., 2019. Association of exposure to phthalates with cardiometabolic risk factors in children and adolescents: A systematic review and meta-analysis. Environmental Science and Pollution Research, 26(35), pp.35670–35686.

Gong, C.-B., Wei, Y.-B., Chen, M.-J., Liu, L.-T, Chow, C.-F., and Tang, Q., 2018. Double Imprinted Photoresponsive Polymer for Simultaneous Detection of Phthalate Esters in Plastics. European polymer journal, 108, pp.295–303.

Gore, A. C., Chappell, V. A., Fenton, S. E., Flaws, J. A., Nadal, A., and Prins, G. S., 2015. EDC-2: the Endocrine Society's Second Scientific Statement on Endocrine-Disrupting Chemicals. Endocrine Reviews, 36, pp.E1–E150.

Giuliani, A., Zuccarini, M., Cichelli, A., Khan, H., and Reale, M., 2020. Critical Review on the Presence of Phthalates in Food and Evidence of Their Biological Impact. The New International Journal of Environmental Research and Public Health, 17 (16), 5655.

Henkel, C., Hüffer, T., and Hofmann, T., 2019. The Leaching of Phthalates from PVC Can Be Determined with an Infinite Sink Approach. MethodsX, 6, pp.2729–2734.

Heo, H., Choi, M. J., Park, J., Nam, T., and Cho, J., 2020. Anthropogenic Occurrence of Phthalate Esters in Beach Seawater in the Southeast Coast Region, South Korea. Water Journal 12 (122), pp.1–13.

Howdeshell, K. L., Rider, C. V., Wilson, V. S., and Gray, L. E., 2008. Mechanisms of Action of Phthalate Esters, Individually and in Combination, to Induce Abnormal Reproductive Development in Male Laboratory Rats. Environmental research, 108 (2), pp.168–176.

He, W., Qin, N., Kong, X., Liu, W., He, Q., Ouyang, H., Yang, C., Jiang, Y., Wang, Q., Yang, B., and Xu, F., 2013. Spatio-temporal distributions and the ecological and health risks of phthalate esters (PAEs) in the surface water of a large, shallow Chinese lake. Science of the Total Environment, 461–462, pp.672–680.

He, M.-J., Lu, J.-F., Wang, j., Wei, S.-Q., and Hageman, K. J., 2020. Phthalate Esters in Biota, Air and Water in an Agricultural Area of Western China, with Emphasis on Bioaccumulation and Human Exposure. Science of the Total Environment, 698, 134264.

Høyer, B. B., Lenters, V., Giwercman, A., Jönsson, B. A. G., Toft, G., Hougaard, K. S., Bonde, J. P. E., and Specht, I. O., 2018. Impact of Di-2-Ethylhexyl Phthalate Metabolites on Male Reproductive Function: A Systematic Review of Human Evidence. Current Environmental Health Reports, 5(1), pp.20–33.

Huang, X., Cui, H., and Duan, W., 2020. Ecotoxicity of Chlorpyrifos to Aquatic Organisms: a Review. Ecotoxicology and Environmental, Safety 200, 110731.

Jeddi, M. Z., Janani, L., Memari, A. H., Akhondzadeh, S., and yunesian, M., 2016. The role of phthalate esters in autism development: A systematic review. Environmental Research, 151, pp.493–504.

Ji, Y.,Wang, F., Zhang, L., Shan, C., Bai, Z., Sun, Z., 2014. A Comprehensive Assessment of Human Exposure to Phthalates from Environmental media and Food in Tianjin, China. Journal of Hazardous Materials, 279, pp.133–140.

Jiang, M., Li, Y., Zhang, B., Zhou, A., Zhu, Y., Li, J., 2018. Urinary Concentrations of Phthalate Metabolites Associated with Changes in Clinical Hemostatic and Hematologic Parameters in Pregnant Women. Environment International 120, pp.34–42.

Koniecki, D., Wang, R., Moody, R. P., and Zhu, J., 2011. Phthalates in Cosmetic and Personal Care Products: Concentrations and Possible Dermal Exposure. Environmental Research, 111 (3), pp.329–336.

Kingsley, O., and Witthayawirasak, B., 2020. Occurrence, Ecological and Health Risk Assessment of Phthalate Esters in Surface Water of U-Tapao Canal, Southern, Thailand. Toxics, 8(3), 58.

Le, T. M., Nguyen, H. M. N., Nguyen, V. K., Nguyen, A. V., Vu, N. D., Yen, N. T. H., Hoang, A. Q., Minh, T. B., Kannan, K., and Tran, T. M., 2021. Profiles of phthalic acid esters (PAEs) in bottled water, tap water, lake water, and wastewater samples collected from Hanoi, Vietnam. Science of the Total Environment, 788, 147831.

Lee, D.-W., Kim, M.-S., Lim, Y.-H., Lee, N., and Hong, Y.-C., 2018. Prenatal and postnatal exposure to di-(2-ethylhexyl) phthalate and neurodevelopmental outcomes: A systematic review and meta-analysis. Environmental Research, 167, pp.558–566.

Lee, Y.-M., Lee, J.-E., Choe, W., Kim, T., Lee, J.-Y., Kho, Y., Choi, K., and Zoh, K.-D., 2019. Distribution of phthalate esters in air, water, sediments, and fish in the Asan Lake of Korea. Environment International, 126, pp.635–643.

Li, M.-C., Chen, C.-H., and Guo, Y. L., 2017. Phthalate esters and childhood asthma: A systematic review and congener-specific meta-analysis. Environmental Pollution (Barking, Essex: 1987), 229, pp.655–660.

Li, R., Liang, J., Duan, H., and Gong, Z., 2017. Spatial distribution and seasonal variation of phthalate esters in the Jiulong River estuary, Southeast China. Marine Pollution Bulletin, 122(1), pp.38–46.

Li, T., Yin, P., Zhao, L., Wang, G., Yu, Q. J., Li, H., and Duan, S., 2014. Spatial–temporal distribution of phthalate esters from riverine outlets of Pearl River Delta in China. Water Science and Technology, 71(2), pp.183–190.

Li, J., Wang, Z.,Wang, Q., Guo, L., Wang, and C.,Wang, Z., 2021. Construction of Hypercrosslinked Polymers for High-Performance Solid Phase Microextraction of Phthalate Esters from Water Samples. Journal of Chromatography A, 1641, 461972.

Ling, W., Gui-Bin, C., Ya-Qi, H., Bin, Y., Wei, W., and Da-Zhong, S., 2007. Cloud point Extraction Coupled with HPLC-UV for the Determination of Phthalate Esters in Environmental Water Samples. Journal of Environmental Sciences, 19, pp.874–878.

Liu, X., Shi, J., Bo, T., Li, H., and Crittenden, J. C., 2015. Occurrence and risk assessment of selected phthalates in drinking water from waterworks in China. Environmental Science and Pollution Research, 22(14), pp.10690–10698.

Loraine, G. A., and Pettigrove, M. E., 2006. Seasonal Variations in Concentrations of Pharmaceuticals and Personal Care Products in Drinking Water and Reclaimed Wastewater in Southern California. Environmental Science and Technology, 40(3), pp.687–695.

Luo, X., Shu, S., Feng, H., Zou, H., and Zhang, Y., 2021. Seasonal distribution and ecological risks of phthalic acid esters in surface water of Taihu Lake, China. Science of the Total Environment, 768, 144517.

Mackintosh, C. E., Maldonado, J. A., Ikonomou, M. G., and Gobas, F. A. P. C., 2006. Sorption of Phthalate Esters and PCBs in a Marine Ecosystem. Environmental Science and Technology, 40(11), pp.3481–3488.

Marie, C., Vendittelli, F., and Sauvant-Rochat, M.-P., 2015. Obstetrical outcomes and biomarkers to assess exposure to phthalates: A review. Environment International, 83, pp.116–136.

Malarvannan, G., Onghena, M., Verstraete, S., van Puffelen, E., Jacobs, A., and Vanhorebeek, I., 2019. Phthalate and Alternative Plasticizers in Indwelling Medical Devices in Pediatric Intensive Care Units. Journal of Hazardous Materials, 363, pp.64–72.

Nantaba, F., Palm, W.-U., Wasswa, J., Bouwman, H., Kylin, H., and Kümmerer, K., 2021. Temporal dynamics and ecotoxicological risk assessment of personal care products, phthalate ester plasticizers, and organophosphorus flame retardants in water from Lake Victoria, Uganda. Chemosphere, 262, 127716.

Navaranjan, G., Takaro, T. K., Wheeler, A. J., Diamond, M. L., Shu, H., and Azad,M. B., 2019. Early Life Exposure to Phthalates in the Canadian

Healthy Infant Longitudinal Development (CHILD) Study: a Multi-City Birth Cohort. Journal of Exposure Science and Environmental Epidemiology 30 (1), pp.70–85.

Net, S., Dumoulin, D., EL-OSMANI, R., Rabodonirina, S., and Ouddane, B., 2014. Case study of PAHs, Me-PAHs, PCBs, Phthalates and pesticides contamination in the Somme River water, France. International Journal of Environmental Research, 8(4), pp.1159-1170.

Net, S., Delmont, A., Sempéré, R., Paluselli, A., and Ouddane, B., 2015. Reliable Quantification of Phthalates in Environmental Matrices (Air, Water, Sludge, Sediment and Soil): A Review. Science of the Total Environment, 515-516, pp.162–180.

Oehlmann, J., Schulte-Oehlmann, U., Kloas, W., Jagnytsch, O., Lutz, I., and Kusk, K. O., 2009. A Critical Analysis of the Biological Impacts of Plasticizers on Wildlife. Philosophical Transactions of the Royal Society B, 364 (1526), pp.2047–2062.

Okpara, K. E., Phoungthong, K., Agbozu, I., Edwin-Isotu, E., and Techato, K., 2022. Phthalate Esters in Tap Water, Southern Thailand: Daily Exposure and Cumulative Health Risk in Infants, Lactating Mothers, Pregnant and Nonpregnant Women. International Journal of Environmental Research and Public Health, 19(4), 2187.

Olujimi, O. O., Aroyeun, O. A., Akinhanmi, T. F., and Arowolo, T. A., 2017. Occurrence, removal and health risk assessment of phthalate esters in the process streams of two different wastewater treatment plants in Lagos and Ogun States, Nigeria. Environmental Monitoring and Assessment, 189(7), 345.

Paluselli, A., Aminot, Y., Galgani, F., Net, S., and sempere, R., 2018. Occurrence of phthalate acid esters (PAEs) in the northwestern Mediterranean Sea and the Rhone River. Progress in Oceanography, 163, pp.221–231.

Poursafa, P., Ataei, E., and Kelishadi, R., 2015. A systematic review on the effects of environmental exposure to some organohalogens and phthalates on early puberty. Journal of Research in Medical Sciences: The Official Journal of Isfahan University of Medical Sciences, 20(6), pp.613–618.

Salaudeen, T., Okoh, O., Agunbiade, F., and Okoh, A., 2018. Fate and impact of phthalates in activated sludge treated municipal wastewater on the water bodies in the Eastern Cape, South Africa. Chemosphere, 203, pp.336–344.

Santhi, V. A., and Mustafa, A. M., 2013. Assessment of organochlorine pesticides and plasticisers in the Selangor River basin and possible pollution sources. Environmental Monitoring and Assessment, 185(2), pp.1541–1554.

Selvaraj, K. K., Sundaramoorthy, G., Ravichandran, P. K., Girijan, G. K., Sampath, S., and Ramaswamy, B. R., 2015. Phthalate esters in water and sediments of the Kaveri River, India: Environmental levels and ecotoxicological evaluations. Environmental Geochemistry and Health, 37(1), pp.83–96.

Seyoum, A., and Pradhan, A., 2019. Effect of Phthalates on Development, Reproduction, Fat Metabolism and Lifespan in Daphnia magna. Science of the Total Environment, 654, pp.969–977.

Sun, J., Wu, X., and Gan, J., 2015. Uptake and Metabolism of Phthalate Esters by Edible Plants. Environmental Science and Technology, 49, pp.8471–8478.

Shi, W., Deng, D., Wang, Y., Hu, G., Guo, J., Zhang, X., Wang, X., Giesy, J. P., Yu, H., and Wang, Z., 2016. Causes of endocrine disrupting potencies in surface water in East China. Chemosphere, 144, pp.1435–1442.

Sibali, L. L., Okonkwo, J. O., and McCrindle, R. I., 2013. Determination of selected phthalate esters compounds in water and sediments by capillary gas chromatography and flame ionization detector. Journal of Environmental Science and Health Part A, 48(11), pp.1365–1377.

Stales C.A., Peterson D.R., Parkerton T.F., Adams W.J., 1997. The environmental fate of phthalate esters: a literature review. Chemosphere, 35, pp.667–749.

Staples, C. A., Adams, W. J., Parkerton, T. F., Gorsuch, J.W., Biddinger, G. R., and Reinert, K. H., 1997. Aquatic Toxicity of Eighteen Phthalate Esters. Environmental Toxicology and Chemistry, 16 (5), pp.875–891.

Talia, C., Connolly, L., and Fowler, P. A., 2021. The Insulin-like Growth Factor System: A Target for Endocrine Disruptors? Environment International 147, 106311.

Tran, B. C., Teil, M. J., Blanchard, M., Alliot, F., and Chevreuil, M., 2015. BPA and phthalate fate in a sewage network and an elementary river of France. Influence of hydroclimatic conditions. Chemosphere, 119, pp.43–51.

UNEP, 2020. An Assessment Report on Issues of Concern: Chemicals and Waste Issues Posing Risks to Human Health and the Environment. Available at: https://wedocs.unep.org/bitstream/handle/20.500.11822/33807/ARIC.pdf?sequence=1&isAllowed=y [Accessed 19 April 2022].

USCDC, 2018. National Health and Nutrition Examination Survey 2015-2016 Data Documentation, Codebook, and Frequencies. Phthalates and Plasticizers Metabolites - Urine (PHTHTE_I). Available at: https://wwwn.cdc.gov/Nchs/Nhanes/2015-2016/PHTHTE_I.htm. [Accessed 19 April 2022]

USEPA, 2018. America's Children and the Environment. Appendix A: Data Tables – Biomonitoring – Phthalates. Available at: https://www.epa.gov/americaschildrenenvironment/data-tables-biomonitoring-phthalates. [Accessed 21 April 2022].

USEPA, 2021. Bis (2-ethylhexyl) Phthalate (DEHP). Available at: https://www.epa.gov/sites/production/files/2016-09/documents/bis-2-ethylhexyl-phthalate.pdf. [Accessed 11 April 2022].

Vabre, P., Gatimel, N., Moreau, J., Gayrard, V., Picard-Hagen, N., Parinaud, J., and Leandri, R. D., 2017. Environmental pollutants, a possible etiology for premature ovarian insufficiency: A narrative review of animal and

human data. Environmental Health: A Global Access Science Source, 16(1), 37.

Vats, S., Singh, R. K., and Tyagi, P., 2013. Phthalates - A Priority Pollutant. International Journal of Advanced Biological, 3 (1), pp.1–8.

Wang, C., Huang, P., Qiu, C., Li, J., Hu, S., Sun, L., Bai, Y., Gao, F., Li, C., Liu, N., Wang, D., and Wang, S., 2021. Occurrence, migration and health risk of phthalates in tap water, barreled water and bottled water in Tianjin, China. Journal of Hazardous Materials, 408, 124891.

Wang, Y., and Qian, H., 2021. Phthalates and Their Impacts on Human Health. Healthcare, 9(5), 603.

Wang, Y., Zhu, H., and Kannan, K., 2019. A Review of Biomonitoring of Phthalate Exposures. Toxics, 7 (2), 21.

Wang, W., Leung, A. O. W., Chu, L. H., and Wong, M. H., 2018. Phthalates Contamination in China: Status, Trends and Human Exposure-With an Emphasis on Oral Intake. Environmental Pollution, 238, pp.771–782.

Weizhen, Z., Xiaowei, Z., Peng, G., Ning, W., Zini, L., and Jian, H., 2020. Distribution and Risk Assessment of Phthalates in Water and Sediment of the Pearl River Delta. Environmental Science and Pollution Research, 27, pp.12550–12565.

Xie, F., Chen, X., Weng, S., Xia, T., Sun, X., and Luo, T., 2019. Effects of Two Environmental Endocrine Disruptors Di-n-butyl Phthalate (DBP) and Mono- N-Butyl Phthalate (MBP) on Human Sperm Functions in Vitro. Reproductive Toxicology, 83, pp.1–7.

Xu, Y., Minhazul, K. A. H. M., Wang, X., Liu, X., Li, X., and Meng, Q., 2020. Biodegradation of Phthalate Esters by Paracoccus Kondratievae BJQ0001 Isolated from Jiuqu (Baijiu Fermentation Starter) and Identification of the Ester Bond Hydrolysis Enzyme. Environmental Pollution, 263, 114506.

Yang, Q., Huang, X., Wen, Z., Shang, Y., Wang, X., Fang, C., and Song, K., 2021. Evaluating the spatial distribution and source of phthalate esters in the surface water of Xingkai Lake, China during summer. Journal of Great Lakes Research, 47(2), pp.437–446.

Yang, J., Li, Y., Wang, Y., Ruan, J., Zhang, J., and Sun, C., 2015. Recent Advances in Analysis of Phthalate Esters in Foods. Trends in Analytical Chemistry, 72, pp.10–26.

Yin, S., Yang, Y., Yang, D., Li, Y., Jiang, Y., Wu, L., and Sun, C., 2019. Determination of 11 Phthalate Esters in Beverages by Magnetic Solid-Phase Extraction Combined with High-Performance Liquid Chromatography. Journal of AOAC International, 102(5), pp.1624–1631.

Yousefi, Z., Ala, A., Babanezhad, E., and Ali Mohammadpour, R., 2019. Evaluation of exposure to phthalate esters through the use of various brands of drinking water bottled in polyethylene terephthalate (PET) containers under different storage conditions. Environmental Health Engineering and Management, 6(4), pp.247–255.

You, H. H., and Song, G., 2021. Review of Endocrine Disruptors on Male and Female Reproductive Systems. Comp. Biochemistry and Physiology, Part C, 244, 109002.

Zhang, Y., Jiao, Y., Li, Z., Tao, Y., and Yang, Y., 2021. Hazards of Phthalates (PAEs) Exposure: A Review of Aquatic Animal Toxicology Studies. Science of the Total Environment, 771, 145418.

Zhang, B., Zhang, T., Duan, Y., Zhao, Z., Huang, X., and Bai, X., 2019. Human Exposure to Phthalate Esters Associated with E-Waste Dismantling: Exposure Levels, Sources, and Risk Assessment. Environment International, 124, pp.1–9.

Zhao, X., Shen, J., Zhang, H., Li, X., Chen, Z., and Wang, X., 2020. The occurrence and spatial distribution of phthalate esters (PAEs) in the Lanzhou section of the Yellow River. Environmental Science and Pollution Research, 27(16), pp.19724–19735.

Zheng, X., Zhang, B.-T., and Teng, Y., 2014. Distribution of phthalate acid esters in lakes of Beijing and its relationship with anthropogenic activities. Science of the Total Environment, 476–477, pp.107–113.

Zhu, Q., Jia, J., Zhang, K., Zhang, H., and Liao, C., 2019. Spatial Distribution and Mass Loading of Phthalate Esters in Wastewater Treatment Plants in China: An Assessment of Human Exposure. Science of the Total Environment, 656, pp.862–869.

Zoeller, R. T., Bergman, A., Becher, G., Bjerregaard, P., Bornman, R., and Brandt, I., 2014. A Path Forward in the Debate over Health Impacts of Endocrine Disrupting Chemicals. Journal of Environmental Health, 13, 118.

Left Blank Intentionally

Smart Farming using AI towards Bangladeshi Agriculture

Mohammed Sowket Ali[1], Abu Muid Md. Raafee[1], Abu Saleh Musa Mia[1]

Abstract

Agriculture is the largest employment sector in Bangladesh. Which has a great impact on the economy, employment generation, poverty alleviation, human resources development, and food circuity of Bangladesh. Traditional Farming uses age-old agriculture equipment whereas Smart Farming is driven by modern latest technology-intensive farming methods. This smart farming system is cost-effective and middle-class farmers can use it in the farm field. This article reviews data-driven smart farming solutions incorporating Artificial Intelligence (AI), Internet of Things (IoT), Robotics and the management involved in the implementation of these smart farming systems in low-income countries like Bangladeshi Agriculture.

Keywords: Agriculture, Precision Agriculture, Artificial Intelligence, Internet of Things, Robotics.

1.1 Introduction

The population of Bangladesh has almost doubled since the 1890s, reaching approximately 161 million people in 2016 which causes food demand to increase briskly. While the land areas for cultivation are quite limited makes it challenging to produce enough food, we also have lots of other farming problems. After the adoption of modern agriculture, mechanization and the green revolution we've reached the digital revolution which takes traditional farming to the next level. The concept of smart Farming refers to the management of agricultural operations with the assistance of contemporary and cutting-edge technological solutions, with the goals of enhancing product quality, increasing output quantity, and decreasing the amount of manual labor required. Even though there is a nuanced distinction between the meanings of Digital Farming, Smart Farming, and Precision Agriculture, many people use them interchangeably.

[1,2,3] Department of Computer Science and Engineering, Bangladesh Army University of Science and Technology, Saidpur, Bangladesh, Email: sowket@gmail.com[1], ammrafie@gmail.com[2], abusalehcse.ru@gmail.com[3].

Farmers generally visit the field to inspect the crop's status and make decisions based on their knowledge and experience. This method is no longer viable since some fields are too vast to be adequately handled, resulting in inefficient resource consumption. Although some farmers have long-term expertise obtained via spending a lot of time working in the field, new technologies can offer a systematic approach for identifying unexpected issues difficult to find by visual examination on periodic tests. Since the global population continues to live a more urban lifestyle, the labor shortage in the farm sector is also becoming an active problem. We also have to deal with challenges such as pest management & agricultural waste alongside both of these. As a result, traditional framing strategies are inefficient to satisfy growing demand and will not take us to agriculture that is productive and sustainable.

Precision agriculture improvements enhanced operating revenue and net return, according to a U.S. Department of Agriculture (USDA) study published in October 2016 (David 2016). There have been significant benefits for farms that chose on being technological in any way, such as reducing effort and conserving resources, increasing productivity or lowering costs with limited manpower, and producing superior food while ensuring environmental sustainability (Díez C., 2017).

The challenge in retrieving crop data using these new technologies is coming up with something consistent and appropriate, whether it is images or numbers because implementation and the information itself are useless. Improved farmer education and training, knowledge exchange, improved access to financial capital, and rising market demand for natural foods are all significant aspects to consider when adopting these technologies (Grand View Research, 2019). Since these technologies are information-based and contain automated services, it's important to remember that the benefits we obtain from them are based on scale economies across each individual farm, as profitability rises as a farm's size grow.

1.2 Agriculture in Bangladesh

1.2.1 Role of Agriculture in Bangladeshi Economic

Agriculture is very essential in Bangladeshi economies where agriculture adds value as a percentage of gross domestic product (GDP). As shown in Figure 1, agriculture accounted for 12.68 percent of Bangladesh's gross domestic product in 2019.

1.2.2 Agriculture Sector of Bangladesh

Crops, livestock, fisheries, and forests are the four key components of the agriculture sector, as shown in Figure 2. Among these, crops make up to 55% of the sector with the most important crops being Aus/Amon/Boro rice paddies, jute,

potato, and wheat (Bangladesh Bureau of Statistics (BBS), 2017). Bangladesh is the world's second-largest jute producer after India. The sector currently makes up 15% of GDP and employs 43% of the workforce (BBS, 2018).

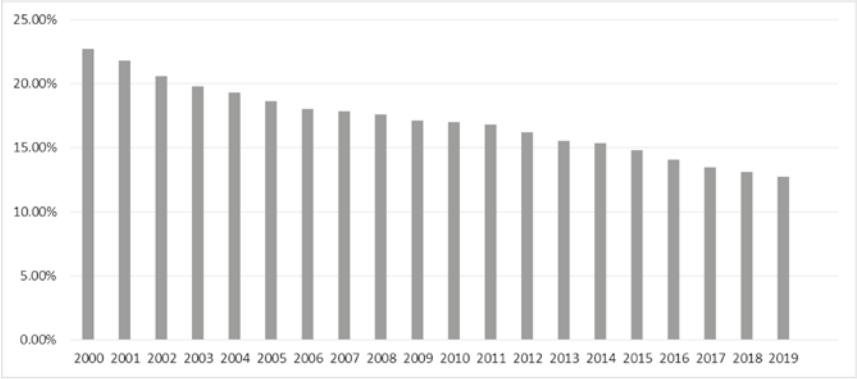


Figure 1: GDP Share of Agriculture

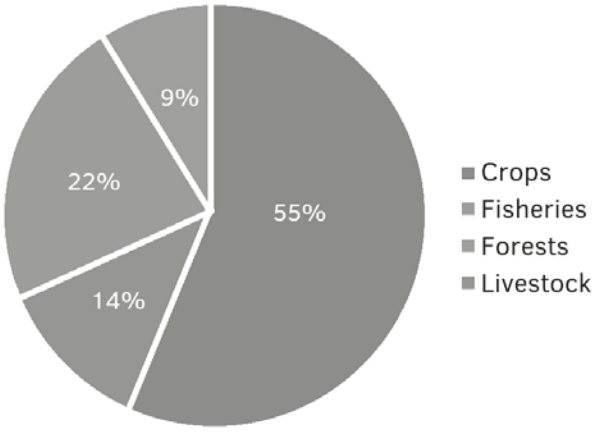


Figure 2: Agriculture Sector Breakdown

Bangladesh's agriculture sector is mainly focused on cereal production. Among cereal grain crops, the primary position is occupied by rice with about 80 percent of the total arable land dedicated to rice cultivation. As a result, rice is the main crop in Bangladesh's agriculture industry, as well as a crucial part of Bangladeshi diets. In Figure 3, it is projected that Bangladesh will be able to supply its own cereal grain with up to 50% of its food grain requirements by 2020.

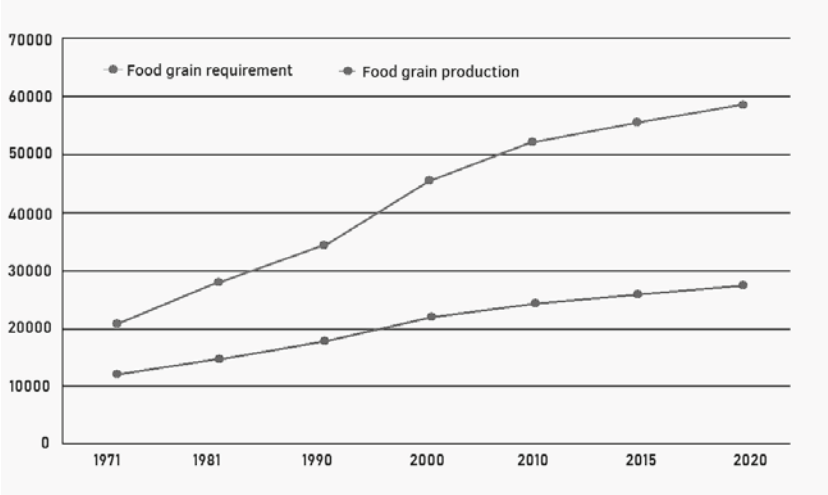


Figure 3: Food grain production and requirement

1.2.3 Present State of Smart Farming in Bangladesh

Technology is like a tidal wave that is not only spreading all over the world but also greatly influencing Bangladesh. In January 2017, The Bangabandhu Sheikh Mujibur Rahman Agricultural University (BSMRAU) started implementing the "E-Village" project while iSoftStone, a joint venture of Chinese Huawei Technologies, who are providing technical support. This device collects real-time data on moisture, air temperature, and leaf health using a number of basic sensors or probes and sends it to the smartphone using Bluetooth in seconds. The online system for receiving real-time support from experienced agriculturalists of BSMRAU makes this much more beneficial.

In April of 2018, Advanced Chemical Industries (ACI) Agribusiness released an android app named 'Fosholi' that serves as the first step toward a comprehensive agriculture platform in Bangladesh. This app can assist farmers by delivering necessary information by coordinating and analyzing ultramodern satellite imagery, advanced agricultural methods, innovative agriculture technology, and climate information.

In 2019 January, Grameenphone introduced the country's first IoT-based digital livestock management solution ‘Digi-Cow’ for livestock farmers. This solution consists of a Base solution and individual Smart tags. Benefits include tracking

body temperature and hormonal changes, determining the right time of insemination, and being informed about cow's illness 48 hours before head. After the reader receives signals sent from the smart tag, it sends signals to the control unit, where its data is inspected and transmitted in real-time as a notification over the internet.

A smart irrigation system has been implemented with an Arduino Uno, micro-controller, sensor integration and water pump. The sensor, which is located in about 15 locations throughout the University of Chittagong, Bangladesh, is used to activate a pump depending on data from the integrated sensor. This process records moisture levels and, as a result, automated the control of the motor consequently (Iqbal A., 2020).

Since 2011, the e-Agriculture pilot project has been carried out by the local administration of Bhatbour Block, Dhighi Union, under Sadar Upazila, in Minikganj District. Data was gathered by the authors between September 1 and September 30, 2015, and they describe the difficulties Bangladeshi farmers encountered when using e-Agriculture services. (M. Rezwan I., 2016)

Juthi Kundu et al. presented a web portal in 2017 that offers an integrated solution to all agricultural issues and confirms the social, economic, and agricultural development of Bangladesh. This website will considerably aid in Bangladesh's growth by eliminating poverty and opening up opportunities for self-employment for the vast majority of the populace.

2. Technologies Used in Advanced Agriculture

2.1 IoT and Big Data

Farmers can use IoT devices to collect environmental metrics and make decisions based on the information. In agriculture, the IoT refers to the use of sensors and other technologies to convert any agricultural feature or activity into data. By 2050, advances in these technologies are expected to boost agricultural productivity by up to 70% (Joe M., 2016). Increased crop output and lower prices are two important benefits of using IoT (Preston G., 2018). The idea of Big Data is popular across a wide spectrum of enterprises in today's technology-driven culture. To collect operational information from bulk data, an automated technique must be applied to the continually growing amount of data collected to manage fields. According to Kunisch, big data looks to be significant in some agricultural scenarios, depending on the farm and the level of technology utilized [Martin Kunisch, 2016). The Consortium of International Agricultural Research Centers

(CGIAR), Montpellier, France has developed a Big Data Platform for Agriculture in order to employ Big Data methods to address agriculture growth concerns more efficiently, swiftly, and effectively much quicker and at a greater level than before (CIAT and IFPRI, 2016).

2.2 Robotics and AI

These days world has evolved from a rural population including a large number of farmers to a city-based system. As a result, farms face the problem of a lack of staff. Agricultural robotics combining Artificial Intelligence (AI) characteristics are one approach to cope with this lack of staff. Robotic agricultural applications are increasing rapidly (Shamshiri R.R, 2018), providing Smart Farming with promising solutions to handle labor shortages and a long-term decrease in profitability. Nevertheless, like other technologies, there are significant constraints to deal with in the present early stages. These technologies are still too expensive for many farmers, particularly anyone with a small farm (Lamborelle, A., 2016), since economies of scale cause small-sized farms less lucrative for food producers (Sonka, S., 2014). Robotic technologies are boosting the global demand for agriculture and crop production, as agricultural robotics would be able to execute field activity more accurately than farmers, as reported by Verified Market Intelligence (VMI, 2018). Modern sensing technologies in agriculture will help solve the problem by providing precise information about land, crop condition, including environmental factors, allowing for accurate application with phytosanitary products, reducing the use of herbicides and pesticides, enhancing the efficiency of water use and increasing crop quality and yield (Zhang Y. 2019).

3. Smart Farming: Management

In advanced-technology adopted farms, field management may differ according to the information-based control process regarding developed agriculture (Sáiz R., 2020). This control strategy, which is focused on realistic field observations with intelligent decision-making, takes place with the crop to be managed, using its internal spatiotemporal versatility. The Information-Based Control Process regarding Developed Agriculture follows the sequence: Crop, Platform, Data, Decision, Actuation.

The platform corresponds to the real means through which information is collected. It is the sensors via which element, objective data is gathered. The information in the data is obtained straight from the plant, soil, and atmospheric parameters. The decision stage is concerned with the use of routines and Artificial

Intelligence Algorithms for sorting out incorrect data and assisting its producer in making the best choices. Lastly, actuation is the process of carrying out a decision-making system-controlled action, which is usually conducted by specialized instruments which would accept instructions from a computerized control unit. The loop continues and closes at the crop level since every activity is carried out on the crop.

4.1 Variability Analyzation

As per Searcy (Stephen W. S., 2011), weather affects natural variability in growing seasons, often throughout the year. So, data from many years could be necessary to establish patterns in any variables of interest. As a result, data contribute significantly to the farm management system on a regular basis. Thus, the importance of crop surveillance arises because variability persists, but the producer must also handle the variability in a proper and efficient manner, which is best accomplished by establishing in-field management zones. Subfield regions of similar traits are referred to as management zones which allows field operations to be tailored to each of them, leading to a realistic and cost-efficient Smart Farming method (Miao Y, 2018).

A number of management zones depend on an area's natural variability, field scale, as well as some management variables (Zhang N., 2002). Besides specifying working areas, it's crucial to carefully pick particular parameters that will be monitored within those zones ahead of time. Klassen, S.P., 2014, stated that, while performing characterization of soil variance in the rice field, there could be particular instances where a field's spatial variability is very poor, in that case, a single mapping event is enough.

4.2 Sensor Integrated Platforms

To track crops and extract objective information from them, sensors are the universal devices. Often, these sensors are integrated into some kind of platform. Such platforms might well be attached to the off-road vehicles and perhaps installed on the ground in areas like local weather stations. Although not every variable of importance can currently be calculated non-invasively or at a distance from its target, but certain developments are seeing major advances, like multi-spectral and hyper-spectral image processing.

4.2.1 Artificial Satellites and Aircraft Systems

As field data from artificial satellites has become readily accessible, remote sensing became very important in the development of Smart Farming. Several

reports on satellite sensing applications have been carried out with recent research focusing on the future uses of remote sensing thermal technology as well as the nutritional status of commodity crops (Rudd J.D., 2017). As sensors work near the target, finer observations are possible. With aircraft systems, the gap between the land and aircraft seems to be about 100 meters. Rotary-wing Unmanned Aerial Vehicles (UAVs) are very stable fliers since they can take off and land vertically, but the problem with these is that they are slow and are incapable of covering much distance throughout their battery life. As fixed-wing platforms cover more area with each flight holding greater payloads, they break quite easily after several landings and tend to be very costly (Rudd J.D., 2017). UAVs have the ability to collect data even from difficult areas that traditional machinery cannot stand as opposed to land vehicles. Although, they require expert travel route planning in advance and certain computer vision implementations may necessitate flight during noon to prevent vegetation shadows which causes image data errors. Because they have limited payload capacity it restricts the maximum number of sensors on board and their inability to travel in high winds. This is an important downside of UAVs.

4.2.2 Autonomous Ground Systems

The space between sensors and the target crop decreases to less than two meters as surveillance platforms work from the ground. Environmental factors like heavy sunshine and insufficient lighting are no longer an issue when active sensors are included and real-time activities like weed spraying with advanced identification of pests are possible when on-the-fly processing is used (Lameski, E., 2018). Aravind addressed land robotics for tillage, crop scouting, soil inspection, planting, pest extermination, replanting, removing weeds as well as harvesting, with crop scouting described as a method of continuous field observation to obtain information regarding crop condition, disease occurrence as well as crop growth-inhibiting infestations (Aravind, K.R., 2017).

4.3 Managing Data

Sensors make it possible to obtain data within the field. However, the unique combo of non-invasive technologies and on-the-fly sensing via moving platforms have unlocked the way to vast data gathering, the precursor to big data in agriculture. Excess data, however, is also a serious challenge to deal with as noise can obscure sensitive information. Since mapping is able to distinguish spatial trends as well as homogeneous regions, a very popular method to view agricultural data seems to be in the form of maps. The aim of constructing maps is to acquire

a few management zones with the criteria of importance in order to accurately implement with care. Kriging became a very common interpolation method for delimiting manageable-sized regions to achieve plausible management zones (Buttafuoco, G., 2016). A coordination system must be provided along with the map when creating a map. The Local Tangent Plane (LTP) coordinate system uses Euclidean geometry, enables user-set origin as well as uses the intuitive east-north coordinate frame, making it an ideal alternative for agricultural maps.

Using Geographic Information Systems (GIS) is a common way of managing field data displayed on maps. Every type of georeferenced data can be stored, analyzed, manipulated, and mapped using these tools. For Precision Agriculture applications (Zhang N., 2001), a particular GIS system named Field-level Geographic Information System (FIS) has been created. The FMIS, which stands for Farm Management Information System is the FIS's modified edition. Its goal is to help the farmers to make the right choices available by minimizing production costs, adhering to agricultural standards, and ensuring good product quality and safety (Fountas S., 2015). Farm management software systems facilitate data acquisition & it's processing and also have the ability to keep track, strategic planning, make decisions, documentation as well as control farm activities (Köksal Ö., 2019), and include essential functionality for keeping records like crop output rates, weather prediction, scheduling of farm tasks, profits and losses, monitoring nutrition level of soil, field mapping, also many complicated capabilities to automate farm and agribusiness field management accounting.

4.4 Making Decisions

People struggle to handle complicated information in circumstances when several field parameters must be evaluated in an attempt to implement effective decisions. AI methods like fuzzy logic, expert systems, neural networks as well as genetic algorithms may aid in these circumstances. AI plays an important part in agriculture, assisting in the interpretation of accessible data, with its modeling as well as reasoning skills. Fuzzy logic, for example, is a method of AI that mimics human thinking by simulating the mechanism of deciding that requires numerous choices than just "true" and "false" options; this strategy utilizes linguistic variables which are well tailored to the complexities of challenges raised by the variety of agricultural decision-making. For potato, corn, and kiwi, Marsili-Libellia and Giusti developed a fuzzy-based Decision Support System (DSS) with rain forecast & soil moisture as input variables (Giusti E., 2015). When various factors are taken into account, DSS can become more stable and efficient, but

certain processes remain contentious since priorities may differ solutions at different times depending upon priorities given by policymakers and other individuals interested in the process (Kumar A., 2017).

4.5 Applying Decisions

Actuation is described as taking actions on the crop or in relation to it, and it is achieved by taking decisions directly after acquiring information or at a later period. Variable Rate Machines have the ability to perform a wide range of farming activities driven via smart systems (Tobe, 2017). Variable Rate Technology (VRT) has the ability of benefit maximization while lowering environmental pollution when applied to Site-Specific Crop Management (SSCM), since just what is necessary would be implemented (Kweon G., 2013). If management zone delineation methods are used in variable-rate nutrient implementations, Nawar et al. reported that field productivity is improved in every instance and environmental damages were minimized in comparison to conventional uniform-rate applications (Nawar S., 2017). The production of commercial VRT solutions is being led by manufacturers of machinery. Automatic differential harvesting, also known as Variable Rate Harvesting (VRH), is another optimistic method of variable actuation that tries to harvest as per previously established management areas. Aside from performance & usefulness, the expense is an important factor to remember when deciding whether or not to implement such technology.

5. Getting Value from AI in Agriculture

Numerous modern logics and approaches have been developed and uncovered as a result of AI development, making the problem-solving process easier. Fuzzy logic, Expert Systems, Neuro-fuzzy logic and Artificial Neural Networks (ANN) are examples of those approaches. ANN is the most generally adopted and frequently used approach for research, across all of these. ANN was created from the same idea about how the human brain works in the mind. ANN is a task-oriented approach that informs the system to function based on an in-built task instead of a computationally programmed task. Due to its superiority over traditional systems, ANNs have been implemented several times throughout the agricultural industry. The primary advantage of neural networks is that, relying on parallel logic, they can predict and forecast. Neural networks can be trained rather than thoroughly programmed. For predicting variables of water supply, Maier & Dandy (Maier, H.R., 2000) had been using neural networks. ANN was used by Gliever and Slaughter (Gliever, C., 2001) to distinguish weeds from crops. Song & He combined ANNs and expert systems to predict crop nutrition levels.

Whenever it gets down to choosing a predicting strategy, neural networks still come out on top (Song, H., He, Y., 2005). If a stable collection of variables is fed to neural networks, they are able to predict complex mappings. To fix the effect of frost development in the fields of Sicily, Mort & Robinson established a neural network-based prediction model (Robinson C. and Mort, N., 1997).

Using substantially less meteorological data, Arif utilized two ANN models to measure moisture levels in paddy fields. The measured and calculated soil moisture values were then used to corroborate and validate these models. To get the estimate Evapotranspiration (ET), the first ANN model was created. The minimum, average, and maximum temperatures of the air were used. Solar, precipitation, as well as air temperature information were collected for the second model. Each of these models generated precise and consistent soil moisture estimates in paddy fields with the least amount of meteorological data, labor, and time. Hinnell discusses neuro drip irrigation systems, in which ANNs were used to model the spatial distribution of water within the subsurface (Hinnell A.C., 2010). Water distribution throughout the lower level of the soil is important for the drip irrigation system to work properly. Here, the prediction made by ANNs is beneficial for the user, resulting in a quick decision-making phase in turn. After the soil has been infiltrated with water from the emitter on the land's surface, ANN models generate wetting trends (first and second). As a result, the ANN model presents the user with continuous trends. In addition, researchers created a model to analyze the maize crop's yield. A Multi-Layered Feedforward ANN (MLFANN) has been used here. Learning algorithms like Gradient Descent Algorithms (GDA) & Conjugate Gradient Descent Algorithm (CGDA) are used to power such networks. Both of the algorithms were written & simulated using the Neural Network Toolbox in the Matrix Laboratory (MATLAB) (Singh, R.K., Prajneshu, 2008).

6.1 Benefit of Smart Farming in Bangladeshi Agriculture

As a result of the increase in the country's population, there is a greater demand for agricultural production in Bangladesh. As a consequence of this, sustainable smart farming is absolutely necessary in Bangladesh in order to raise crop production despite the country's limited resources. In this paper we figure out that the smart farming implementation in Bangladesh will result in a number of benefits, including the following Table 1:

Table-1: Benefit of Smart Farming in Bangladeshi Agriculture

SN	Benefits of Smart Farming
1	Producing a greater quantity of crops at a lower cost.
2	Because there is a limited of water, reduce the amount of water that is w asted.
3	Producing more crops on a limited amount of land.
4	Automate the Irrigation System.
5	Can be notified in advance of a variety of alarms, such as temperature a nd air humidity, an inadequate water balance at the time of irrigation, or any unusual activities that take place in the field.
6	Decreases the amount of fertilizer and insecticides that are needed.
7	Agricultural Operations can be managed by remote Control
8	The farmer can also administer and monitor the system using his or her smartphone.
9	Enhancing indoor safety.
10	Reduce physical labor in the field and help to reduce the number of workers as well.
11	Bangladeshi people's lifestyles are becoming more comfortable and getting better living environment.
12	The economy of Bangladesh should be increased.

6.2 Challenges of Smart Farming in Bangladeshi Agriculture

The utilization of e-agricultural services is problematic for about 65.8% of farmers (Md. Rezwan Islam, 2016). According to our analysis, the following challenges are faced by Bangladeshi farmers to implement smart farming.

Table-2: Challenges to Implement Smart Farming in Bangladesh

SL	Challenges
a)	Lack of knowledge of the advantages of smart farming.
b)	Farmers' inability to use mobile-based apps due to a lack of e-literacy.
c)	The cost of using, developing, and implementing Smart Farming
d)	Timely and effective delivery
e)	Inadequate government digital services facilities and centers
f)	Essential Data Centers
g)	Internet accessibility
h)	Supply of electricity
i)	Inadequate IT professionals
j)	Inadequate training
k)	Easily Operable Technology

7. Conclusion

In recent decades, as advances in technology have emerged, particularly access to accurate agricultural data and excellent computer tools to extract the most significant information from it, we can ultimately obtain maximum benefits while being environment friendly. It seems that it's the right moment to take steps toward digital, sustainable agriculture capable of demonstrating the maximum potential of data-driven management in the twenty-first century in order to meet the demands of food production. It can perform jobs quicker and better than humans with the integration of Robotics, IoT and AI, leading to very productive farm management. In order to prevent repeated, physically exhausting, and stressful field operations, these modern solutions based on digital technologies allow farmers to serve as supervisors of ones crops instead of just laborers. However, users, especially young farmers eager to understand and apply advanced technologies to agriculture, must undergo intensive training in order to take full advantage of advanced agriculture, and ensure a generational rebirth.

References

Amlan H., Nahina I., Nahidul H. S., Shuvashis D. and Biplob R., 2021, Smart Farming through Responsible Leadership in Bangladesh: Possibilities, Opportunities, and Beyond, Sustainability 2021, 13,4511, MDPI, pp.1-18.

Aravind, K.R., Raja, P., and Pérez-Ruiz, M., 2017, Task-based agricultural mobile robots in arable farming: A review. Span. J. Agric. Res., 15, pp.1–16.

Arif C., Mizoguchi M., Setiawan B.I. and Doi, R., 2012. Estimation of soil moisture in paddy field using Artificial Neural Networks. International Journal of Advanced Research in Artificial Intelligence. 1 (1), pp.17–21.

BBS 2017, Bangladesh Strategic Plan on Agricultural and Rural Statistics (2016-2030) – Bangladesh Bureau of Statistics(BBS), pp.1-62.

BBS 2018, 45 Years of Agricultural Statistics on Major Crops (Aus, Amon, Boro, Jute, Potato & Wheat). Bangladesh Bureau of Statistics(BBS), January, pp.1-216.

Buttafuoco, G. and Lucà, F., 2016, The Contribution of Geostatistics to Precision Agriculture. Ann. Agric. Crop Sci., 1, pp.1008–1009.

CIAT and IFPRI, 2016. CGIAR Big Data Coordination Platform. Proposal to the CGIAR Fund Council. Available online: https://cgspace.cgiar.org/handle/10947/4303 [Accessed 21/06/2022].

Díez C., 2017. Hacia una agricultura inteligente (Towards Smart Agriculture). Cuaderno de Campo, 60, pp.4–11.

Fountas, S., Carli, G., Sørensen, C.G., Tsiropoulos, Z., Cavalaris, C., Vatsanidou, A., Liakos, B., Canavari, M., Wiebensohn, J.and Tisserye, B. 2015, Farm management information systems: Current situation and future perspectives. Comput. Electron. Agric., 115, pp.40–50.

Giusti, E., Marsili-Libelli and S. 2015, A Fuzzy Decision Support System for irrigation and water conservation in agriculture. Environ. Model. Softw., 63, pp.73–86.

Gliever C. and Slaughter, D.C., 2001. Crop verses Weed recognition with Artificial Neural Networks. American Society of Agricultural and Biological Engineers. Paper Number 013104, pp.1–12.

Grand View Research (GVR), 2019. Precision Farming Market Analysis. Estimates and Trend Analysis; Grand View Research Inc.: San Francisco, CA, USA; pp. 1–58.

Hinnell A.C., Lazarovitch N., Furman A., Poulton M. and Warrick A.W., 2010. Neuro-drip: estimation of subsurface wetting patterns for drip irrigation using neural networks. Irrig. Sci. 28, pp. 535–544.

Iqbal A. and Manoranjan S. A., 2020, IoT based Smart Irrigation System at University of Chittagong, Bangladesh, International Journal of Computer Applications (0975 – 8887), Volume 176 – No. 15, April, pp.39-45.

Joe M. and Junko K., 2016, From Dirt to Data: The Second Green Revolution and the Internet of Things. Deloitte Insights. Issue 18. Available at: https://www2.deloitte.com/insights/us/en/deloitte-review/issue-18/second-greenrevolution-and-internet-of-things.html #end note-sup-9 [Accessed 21/06/2022]

Juthi K., Supriya D., Saleh A. and Sajal H., 2017, Smart E-Agriculture Monitoring System: Case Study of Bangladesh, Proceedings of the 2017 4th International Conference on Advances in Electrical Engineering (ICAEE), 28-30 September, Dhaka, Bangladesh

Khanal, S., Fulton, J., and Shearer, S., 2017, An overview of current and potential applications of thermal remote sensing in precision agriculture. Comput. Electron. Agric., 139, pp.22–32.

Klassen, S.P., Villa, J., Adamchuk, V., and Serraj, R., 2014, Soil mapping for improved phenotyping of drought resistance in lowland rice fields. Field Crops Research, Volume 167, pp.112–118.

Kumar, A., Sah, B., Singh, A.R., Deng, Y., He, X., Kumar, P. and Bansal R.C., 2017, A review of multi criteria decision making (MCDM) towards sustainable renewable energy development. Renew. Sustain. Energy Rev., 69, pp. 596–609.

Köksal, Ö. and Tekinerdogan, B. 2019, Architecture design approach for IoT-based farm management information systems. Precis. Agric., 20, pp. 926–958.

Kweon, G., Lund, E. and Maxton, C. 2013, Soil organic matter and cation-exchange capacity sensing with on-the-go electrical conductivity and optical sensors. Geoderma (Elsevier), Volume 199, pp.80–89.

Lamborelle A., and Fernández Á. L., 2016, Farming 4.0: The Future of Agriculture?, EURACTIV France, Available online: https://www.euractiv.com/section/agriculture-food/infographic/farming-4-0-the-future-of-agriculture/ [Accessed 21/06/2022].

Martin K., 2016. Big Data in Agriculture—Perspectives for a Service Organization. Landtechnik Agricultural Engineering, 71, pp. 1–3.

Maier H.R. and Dandy G.C., 2000. Neural networks for the prediction and forecasting of water resources variables: a review of modeling issues and applications. Environmental Modeling and Software, Volume 15, Issue 1, pp.101–124.

M. Nazmus S. M. , Zenville E., M. A. Razzaque , Marco Zennaro and Antoine Bagula, 2016, Enabling The Internet of Things in Developing Countries: Opportunities and Challenges, 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV), pp.564-569.

Miao Y., Mulla D.J. and Robert P. C., 2018. An integrated approach to site-specific management zone delineation. Front. Agric. Sci. Eng., 5, pp. 432–441.

Nawar, S., Corstanje, R., Halcro, G., Mulla, D. and Mouazen, A.M., 2017, Delineation of Soil Management Zones for Variable-Rate Fertilization: A Review. Advances in Agronomy (Elsevier), Volume 143, pp.175–245.

Olly R. C. and Sarna M., 2021, Prospect of IoT Based Smart Agriculture in Bangladesh-A Review, International Journal of Engineering Research & Technology (IJERT), Vol. 10 Issue 11, November.

P. Lameski, E. Zdravevski, and A. Kulakov, 2018, Review of Automated Weed Control Approaches: An Environmental Impact Perspective. ICT Innovations 2018. Engineering and Life Sciences, pp. 132–147.

Preston G., 2018, Precision Agriculture Yields Higher Profits, Lower Risks. Hewlett Packard Enterprise Company. Available online: https://www.hpe.com/us/en/insights/articles/precision-agriculture-yields-higher-profits-lower-risks-1806.html [Accessed 21/06/2022].

Robinson C. and Mort N., 1997. A neural network system for the protection of citrus crops from frost damage. Comput. Electron. Agric. 16 (3), 177–187.

Rudd, J.D., Roberson, G.T., and Classen, J.J., 2017, Application of satellite, unmanned aircraft system, and ground-based sensor data for precision agriculture: A review. In Proceedings of the 2017 ASABE Annual International Meeting; American Society of Agricultural and Biological Engineers, Spokane, WA, USA, 16–19.

Sáiz R. V. and Rovira M. F., 2020, From Smart Farming towards Agriculture 5.0: A Review on Crop Data Management. Agronomy. 10(2):1-21.

Schimmelpfennig and David. 2016. Farm Profits and Adoption of Precision Agriculture. ERR-217, U.S. Department of Agriculture, Economic Research Service, 217, 1–46.

Shamshiri R.R., Weltzien, C., Hameed, I.A., Yule I.J., Grift, T.E., Balasundram S.K., Pitonakova L., Ahmad, D. and Chowdhary, G., 2018, Research and development in agricultural robotics: A perspective of digital farming. Int. J. Agric. Biol. Eng., 11, 1–14.

Sheikh M. M. R. and M. Rezwan I., 2016, Problems faced by farmers in application of Agriculture in Bangladesh, Journal of Agricultural Economics and Rural Development,Vol. 3(1), pp. 079-084, May.

Singh R.K., and Prajneshu, 2008. Artificial neural network methodology for modelling and forecasting maize crop yield. Agric. Econ. Res. Rev. 21, 5–10.

Song H. and He, Y., 2005. Crop nutrition diagnosis expert system based on artificial neural networks. Third International Conference on Information Technology and Applications (ICITA'05), Sydney, NSW, 1, pp. 357–362.

Sonka S., 2014, Big Data and the Ag Sector: More than Lots of Numbers. Int. Food Agribus. Manag. Rev., 17, 1–20.

Stephen W. S., 2011, Precision Farming: A New Approach to Crop Management. L-5177, Texas Agricultural Extension Service, Available online [Accessed 21/06/2022],https://agrilifecdn.tamu.edu/lubbock/files/2011/10/precisionfarm_1.pdf

Tobe, 2017, The Ultimate Guide to Agricultural Robotics. Robotics Business

Review, January 1, Available at online: https://www.roboticsbusinessreview.com/agriculture/ the_ultimate_guide_to_agricultural_robotics/ [Accessed 21/06/2022].

Verified Market Intelligence (VMI), 2018, Global Agriculture Robots. Market Size, Status and Forecast to 2025; Boonton, NJ, USA, pp. 1–79.

Zhang N., Wang M., and Wang, N., 2002, Precision agriculture—A worldwide overview. Comput. Electron. Agric., 36, pp.113–132.

Zhang N., and Taylor R.K. 2001, Applications of a Field–Level Geographic Information System (FIS) in Precision Agriculture. Appl. Eng. Agric., 17, pp.885–892.

Zhang Y. 2019, The Role of Precision Agriculture Resource, Resource Magazine 26(6), pp. 9-9.

Detection of Fake Job Postings on Online Using Convolutional Neural Network

Md Istakiak Adnan Palash[1]
Arijit Diganto[1]
Osama Nazmul Fatan[1]
Kazi Abu Taher[1]
Md Jaber Al Nahian[1,2]

Abstract

The present era focuses on every aspect of modern civilization that can be handled online, such as internet banking, teaching, safety, and employment, etc. This advancement in technology makes it easy for scammers to make money very quickly by looting people. Fake job advertisements are among the latest scams. When people apply for these fake jobs, they have topay fees and send their personal information to the fraudsters, which results in a scam and losing money. Therefore, in this paper, we have proposed a novel Convolutional Neural Network(CNN) to identify fake job postings efficiently. A publicly available dataset named EMSCAD was used to validate our proposed model. A comparison was also made between our proposed model and several state-of-the-art machine learning algorithms. In our experiments, we found that our proposed model had a greater accuracy than other machine learning algorithms. In addition, this study conducts a critical comparison of our method with the most recent existing studies.

Keywords: Fake Job Posting; COVID-19; Detection; Machine Learning; CNN.

1. Introduction

As the corona virus struck the world, the whole world took a very big blow. Everything stood still for a brief period of time, as everyone was instructed to maintain social distancing. It mostly impacted the job industry; many people lost their jobs. While maintaining the social distance, taking recruitment for companies were very difficult and time consuming. So, everything went online. Many people took the liberty of this situation to scam other people. Coronavirus 2019 (COVID-19) pandemic initially occurred in Wuhan, China, in 2019, and

[1] Department of Information and Communication Technology, Bangladesh University of Professionals, Dhaka, Bangladesh.
[2] Corresponding Author: jaber.nahian@bup.edu.bd

quickly spread throughout the world, posing a serious public threat in every sector of China to everywhere (Atangana & Araz, 2020). The COVID-19 epidemic affected many lives around the world and posed an unprecedented threat not only to public health and food systems, but also to the job sectors. According to a report of the World Bank, due to the pandemic around 68 percent of the people of Dhaka and Chittagong in Bangladesh have lost their job in urban areas (Genoni et al., 2020). According to this survey, the rate of job lost in the capital was 76 percent, while it was 59 percent in the port city. A certain group of people is abusing these situations for their own benefit by creating fake jobs to scam helpless people. Scammers are targeting a large number of people who have lost their jobs or seen their wages decrease as a result due to Covid- 19 outbreak. They are mainly scamming people by work-from-home jobs or remote jobs related scams. Many people are looking for legitimate work-from-home jobs because of lay off, pay cuts, quarantines, and stay-at-home orders. Also, some who have a job are also trying to earn more to pay off debt. Some examples of job scams are work-from-home job scams,job placement service scams, government, and postal job scams, nanny, caregiver, virtual personal assistant job scams, and mystery shopper scams (*Job Scams*, 2020). The Ministry of Liberation War Affairs of Bangladesh, had warned the public in November 2020 about a fraudulent job circular circulated by a private entity named 'Grameen Service Bangladesh Limited' claiming government approval, which is totally inaccurate and motivated (*Govtwarns on fake job circular by Grameen Service Bangladesh Ltd", Govt warns on fake job circular by Grameen Service Bangladesh Ltd*, 2022). Also, high-income economic countries faced this problem. During the lockdown, millions of Americans were out of work since the fake jobs have grown (Reinicke, 2020). According to SAFER jobs, a charity that assists flexible economy workers, during the lockdown, the number of fake job circulars increased by 66% in the UK (Burke, 2020). As per the aforementioned problems, early detection of fake job posts is important. Therefore, in this paper, we have proposed a noble CNN-based classification model to detect fake job postings on different platforms. The main contribution of this work is given below:

a) We used a publicly available dataset to validate our model.

b) Some preprocessing steps have been performed before the data is fed into the proposed CNN model.

c) Finally, we have compared our proposed model's performance with the state-of- the-art models and different machine learning algorithms. The results showed that our proposed model outperformed the existing models.

The following sections make up the remainder of this paper: Section 2 describes the existing works, and Section 3 demonstrates the methodology of the proposed model. Result analysis & discussion is presented in Section 4 and the final section

of this paper concludes the whole work.

2. Related Works

In recent times, especially during the COVID-19 period, people's dependence on online has increased significantly. On the other hand, people are being deceived in various ways by online fraudsters who snatch important information from them. A group of fraudsters is currently cheating on the common people through fake job advertisements, stealing their personal information and money from them. Many researchers have already worked to identify online fake job advertisements. Different machine learning and deep learning algorithms have been applied for predicting fraud jobs.

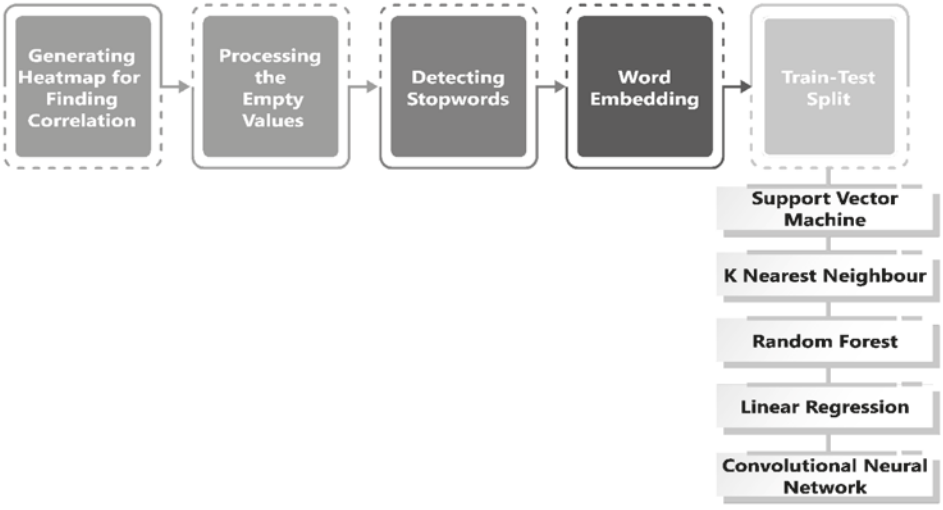


Figure 1: A workflow diagram of the proposed work

Vidros et al. (2017a) first introduced this online fraud. They have also introduced the Employment Scam Aegean Dataset (EMSCAD) publicly. As the dataset is imbalanced, they have taken equal samples from both fraudulent and non-fraudulent jobs. They have described the factors related to fraud and gave a brief overview of the features. They split their experiment into two steps. In the first step, they used bow model and in the second step, they transformed all the data from a balanced dataset into a vector of binary features. Lastly, they compared their two different approaches and their performance. The random Forest algorithm performed better in two cases with an accuracy of 91.22% and 90.56% respectively.

Alghamdi et al. (2019) proposed an intelligent model for the detection of fraudulent and non-fraudulent jobs. They have used SVM for feature selection

and Random Forest for classification purposes and it gives 97.41% accuracy. They conducted their whole experiment with Weka tools and divided their working procedure into three different stages. They are the pre-processing stage, feature selection stage, and a classification stage. Company profile, has the company logo and Industry are the main three features that are extracted from the dataset.

Dutta and Bandyopadhyay (2020) presented a single classifier and ensemble classifier- based model for the detection of fraudulent jobs. They adjusted the parameters for different algorithms. In MLP they used five hidden layers and for K-NN they set the value of K=5. On the other hand, 500 estimators have been used for ensemble classifiers. For the evaluation of the models, they used different methods like Accuracy, F-1 score, Cohen- Kappa Score, etc. Finally, Decision Tree and Random Forest performed better for single classifier and ensemble classifier.

Habiba et al. (2021) used different data mining techniques for the classification of fake jobs. They followed two different procedures for the classification. First, they used conventional machine learning algorithms for the detection of fake job postings and in the second step, they used deep neural networks to the detection of the fake jobs. They used 10-fold cross-validation for training purposes. Their model achieved 99 % accuracy for DNN (fold 9) and 97.7% classification accuracy on average for Deep Neural Network.To reduce the overfitting of the model, they used the dropout layer. Lastly, they showed that Deep Neural Network performs much better than conventional algorithms.

Anita et al. (2021) applied machine learning and deep learning algorithms to differentiate real jobs and fake jobs. They also emphasized on data analysis and data cleaning. The data analysis part gives an insight into the data. In the data cleaning part, they removed the columns which have very large null values. Then they removed all the stop words and combined all textual data together in one column for the further process.

On the other hand, Ranparia et al. (2020) trained their model as a Sequential Neural Network and used the Global Vector model (GloVe) algorithm. For analyzing the pattern of the data, they used Natural Language Processing. In Exploratory Data Analysis they minimize computation, remove noisy data, visualize the data, and improve model accuracy. For the testing purpose, they predict 138 actual jobs which they collect from LinkedIn. Their model predicts 136 job postings as non-fraudulent which indicates that 99.27% of jobs are correctly detected by the model.

Considering the class imbalance problem in mind, Vo et al. (2021) presented an oversampling technique known as Fake Job Description Detection Using Oversampling Techniques for reducing the imbalance of the dataset. They used the

Employment Scam Aegean (ESA) dataset in their experiment. Firstly, they preprocessed their dataset by removing stop words, and tokenizers. Then they extracted features using a bag of words and term frequency-inverse document frequency which converted the feature into a vector. Before training, they used SVMSMOTE which is an oversampling technique. It is applied to balance the imbalance dataset by generating new synthetic samples. After that, they trained their model. They compared both results which they found before applying over-SVMSMOTE and after applying over-SVMSMOTE where Logistic Regression gives 86.60% accuracy and over-SVMSMOTE-LR gives 92.02% accuracy respectively. In both cases, they found different results where after applying over-SVMSMOTE the model performed better, and the accuracy also increased significantly.

Shibly et al. (2021) used the Microsoft Azure Machine Learning Studio platform for their experiment. They compared the performance of two different types of algorithms. They are the Two-Class Boosted Decision Tree algorithm and Two-Class Decision Forest algorithms and evaluate the performance of the algorithms considering the Accuracy, F1 Score, Recall, and Precision.

Lal et al. (2019) proposed an ensemble learning-based model called ORF Detector. They categorized the features of the dataset into three parts: Linguistic, Contextual, and Meta-data. For building their model they used three base classifiers and three ensemble techniques. They used WEKA for the implementation of the algorithms. The proposed model gives 95.4% accuracy and performed better than the baseline classifiers.

Shree et al. (2021) divided their experiment into five modules. They are dataset collection & preprocessing, data visualization, applying classification algorithm, evaluation, result, and analysis. They also applied the feature selection technique, checked for missing values, and visualized the data. After that, they trained and tested their model using three different algorithms where Random Forest Classification performed better than others which gives 99.8% accuracy.

Mehboob and Malik (2021) proposed a fraud detection framework. They used EMSCAD dataset for their experiment. For their experiment model, they followed data preparation, and feature selection and then applied machine learning algorithms. For better investigation of the impact of the fraud, they categorized them into three different features. They are the organization features, job description features, and compensation features. In their proposed model XGBoost performed better than all other algorithms.

3. Methodology

In this section, we have discussed about the working procedure of our research work. The workflow diagram of our research work is shown in Figure 1. The following sub-sections will describe our used dataset, data preprocessing, and

proposed CNN architecture.

3.1 Data collection

We used a dataset published by the University of Aegean which is called the Employment Scam Aegean Dataset (EMSCAD) (Vidros et al., 2017). This is a publicly available dataset, and this dataset is also available on Kaggle. The dataset contains 18 columns describing different types of features of a specific job These columns tell us about the location of the job, which department we are applying to, what the salary would be, company name, job description, what are the requirements for the job, the employment type, and many more. Some of the columns have numeric values such as the columns 'has company logo', 'has question', 'telecommuting', and 'fraudulent'. The values in these columns are either 1 or 0. The rest of the columns have text values. In the dataset there are two types of jobs posted, types are fraudulent and non-fraudulent. Among the total of 17800 job posted in the dataset, number of non-fraudulent jobs were 17014 and 866 were fraudulent jobs. So, the percentage of the jobs were:

Non-fraudulent jobs: (17014/17800) X 100% = 95%

Fraudulent jobs: (866/17800) X 100% = 5%

3.2 Data preprocessing

The dataset is in a CSV format. It has both numeric and text data. We need a tool to read data from the CSV file. For this reason, the best option available is pandas in python. Pandas is a library in the python programming language for manipulating and analyzing data. After that, we looked for NAN values in all the columns. There were no numerical missing values but for the missing text values, we replaced the NAN values with a space(' '). Then we created a variable called text. This contains the concatenated values from the following columns: company profile, description, requirements, and benefits. Whitespace was added before each concatenation point. There is a sentence formed from the above- mentioned columns for each job. We will tokenize each and every sentence.

Now comes the part of processing this data and splitting the data into a test set and train set. Before doing tokenization, removal of stop words was done with the help of NLTK library. Natural Language toolkit or nltk is a library used for preprocessing text data. Stop words in any language are those words which does not add much meaning to a sentence. Even if we discard these words from a sentence, the meaning of the sentence won't change that much.

To split the dataset into train set and test set we will use a library called sklearn. Scikit- learn or sklearn is a machine learning library for python. It has many features. Among all those features we will use train test split. This is a function in model selection for splitting the dataset into two subsets. With the help of this function, we don't have to manually split the dataset. It will split the dataset

randomly. We split the dataset into 75% training data and 25% test data.

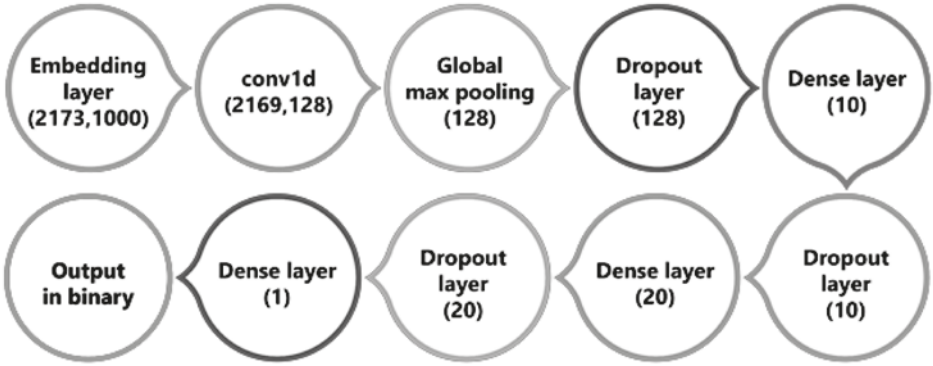


Figure 2: Proposed CNN architecture

After splitting the dataset, we will tokenize the sentences both in the training set and the test set. We will use the Keras tokenization function. After fitting the model, we used a method called texts to sequence both on the train set and the test set. It converts each sentence into a sequence of integers. Only words known by the tokenizer are taken into account. All the sentences aren't of the same length. So the model may emphasize the data with greater length. Hence, to avoid this problem we will add zeros in every sequence. This will make every sequence of equal length. This process was done using the pad sequences method. The maxlen parameter in this method was equal to 2173. This was the maximum length we found of a sequence. So now every sequence is of equal length.

3.3 Proposed CNN Model

We proposed a novel convolutional neural network to detect fake job postings. The proposed CNN is made up of several layers. The first layer is embedding layer which sizeis (2173,1000). Convolution layer is like applying a filter to the input layer to extract the required information. It has filter size 128 and kernel size 5. It passes the extracted features to the next layer, which is max pooling layer. Since the size of the input was very big, wewould need much more time to process the data. So, to reduce the dimensionality we onlyextracted the most prominent features from the previous layer. Hence the output from this layer would be the most notable features of the previous feature map. The next layer is dropout layer, which has the same filter size as the con1d layer. Dropout layers are added to avoid overfitting. The next layer is dense layer, it is a layer of fully connected layer. In this layer, every neuron received input from every neuron of the previous layer. It was done to create diversity in the input type. As in to mix up the extracted features to detectany missing clues. In the dense layers we used relu as the activation function. Another dropout and dense layer

were added on, and then the features were passed on to the outputlayer. The binary classification was done on the output layer, determining whether the job was a fake job or not. On the output layer, sigmoid activation function was used, since the output would be either 0 or 1. Batch size of 50 and 10 epochs were run over the dataset. Figure 2 shows the architecture of the proposed CNN model.

4. Result Analysis and Discussion

In this experiment, we have implemented both the classical machine learning algorithms and Convolutional Neural Network (CNN) for the classification of fake job postings. First, we have implemented machine learning-based model by using Logistic Regression (LR), Random Forest (RF), K-Nearest Neighbours (KNN), and Support Vector Machine (SVM) algorithms. From these algorithms SVM gives the highest accuracy which is 98.45%. Logistic

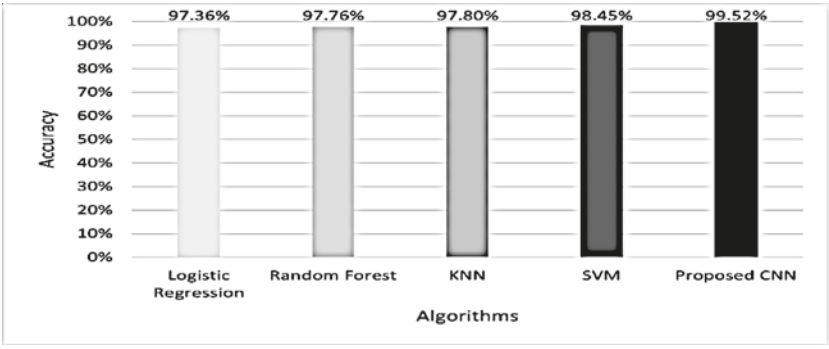


Figure 3: Performance comparison of our proposed CNN model with different state of the art machine learning model.Regression,

Random Forest (RF), and K-Nearest Neighbours (KNN) give an accuracy of 97.36%,97.76%, and 97.80% respectively. A performance comparison of above-mentioned machine learning models with the proposed CNN model is given in Figure 3.

After that, we implemented our proposed model using a Convolutional Neural Network (CNN) and then compared the results. Our proposed model performs better than the other machine learning model in this case with an accuracy of 99.52%. It shows an indication that our proposed CNN model can perform well to detect fake job postings. A performance comparison among existing works (Alghamdi et al., 2019; Habiba et al., 2021; Lal et al., 2019; Ranparia et al., 2020; Vidros et al., 2017a) is shown in Figure 4. The bar

chart shows that our model performs better than the existing studies.

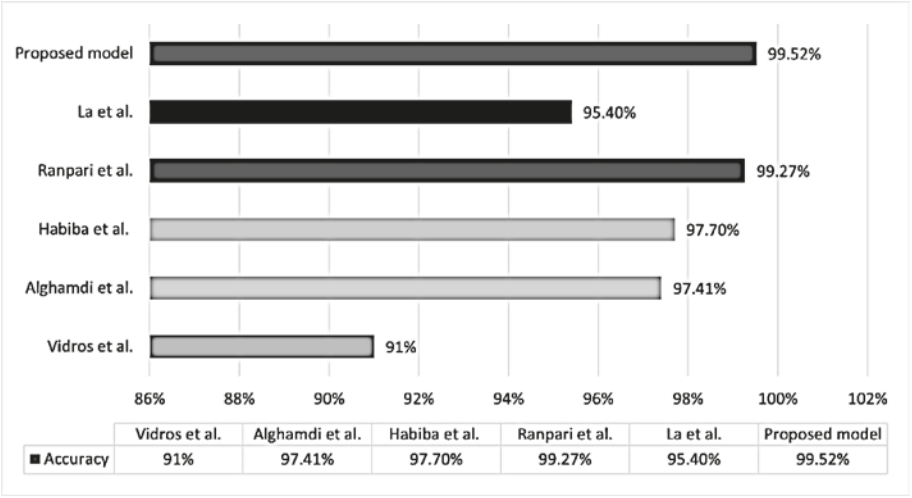


Figure 4: Performance comparison of existing works with our proposed CNN model.

5. Conclusion

Online recruitment is a big part of the recruitment and hiring process for a company or organization. By doing online recruitment they reach out to much more people from different places rather than staying local. During covid-19 situation, many people lost their jobs. In this post-covid situation, people are desperately looking for jobs. Therefore, some people will try to scam others by posting fake jobs online and taking money from them. We built a simple, yet very accurate fake job posts detecting model. We proposed a Convolutional Neural Network (CNN) to detect if a job post is a fake post or a genuine post and we got an accuracy of 99.52%. We worked with the Employment Scam Aegean Dataset. We had a very limited amount of data to work with. We compared our proposed model with different state of the art machine learning models such as SVM, Logistic Regression, K-Nearest Neighbours, and Random Forest. It is showed that our CNN model provided a better result than all of the above. This will help prevent people from being scammed. Also, the model can be further improved with more data from the recent following times.

References

Alghamdi, B., Alharby, F., et al. (2019). An intelligent model for online recruitment fraud detection. Journal of Information Security, 10(03), 155.

Anita, C., Nagarajan, P., Sairam, G. A., Ganesh, P., & Deepak kumar, G. (2021). Fake job detection and analysis using machine learning and deep learning algorithms. Revista Geintec-Gestao Inovacao e Tecnologias, 11(2), 642–650.

Atangana, A., & Araz, S. I˙. (2020). Mathematical model of covid-19 spread in turkeyand south africa: theory, methods, and applications. Advances in Difference Equations, 2020(1), 1–89.

Burke, J. (2020). Covid-19 scam — fake job ads set to rise, warns expert. https:// www.hrgrapevine.com/content/article/2020-09-04-fake-job-ads -set-to-rise-warns-expert. (Accessed: 2022-06-05)

Dutta, S., & Bandyopadhyay, S. K. (2020). Fake job recruitment detection using machine learning approach. International Journal of Engineering Trends and Technology, 68(4), 48–53.

Genoni, M. E., Khan, A. I., Krishnan, N., Palaniswamy, N., & Raza, W. (2020). Losing livelihoods: The labor market impacts of covid-19 in bangladesh. World Bank.

Govt warns on fake job circular by grameen service bangladesh ltd", govt warns on fake job circular by grameen service bangladesh ltd. (2022). https://m.theindependentbd.com/post/255536. (Accessed: 2022-06-05)

Habiba, S. U., Islam, M. K., & Tasnim, F. (2021). A comparative study on fake job post prediction using different data mining techniques. In 2021 2nd international conference on robotics, electrical and signal processing techniques (icrest) (pp. 543–546).

Job scams. (2020). https://consumer.ftc.gov/articles/job-scams. (Accessed: 2022-06-05)

Lal, S., Jiaswal, R., Sardana, N., Verma, A., Kaur, A., & Mourya, R. (2019). Orf detector: ensemble learning based online recruitment fraud detection. In 2019 twelfth interna- tional conference on contemporary computing (ic3) (pp. 1–5).

Mehboob, A., & Malik, M. (2021). Smart fraud detection framework for job recruitments. Arabian Journal for Science and Engineering, 46(4), 3067–3078.

Ranparia, D., Kumari, S., & Sahani, A. (2020). Fake job prediction using sequential net- work. In 2020 ieee 15th international conference on industrial and information systems (iciis) (pp. 339–343).

Reinicke, C. (2020). Job scams have increased as covid-19 put millions of americans out of work.here's how to avoid one. https://www.cnbc.com/2020/10/06/job-scams-have increased-during-the-covid-19-crisis-how-to-one.html. (Accessed:2022-06-05)

Shibly, F., Uzzal, S., & Naleer, H. (2021). Performance comparison of two class boosteddecision tree and two class decision forest algorithms in predicting fake job postings.

Shree, R. A., Nirmala, D., Sweatha, S., & Sneha, S. (2021). Ensemble modeling on job scam detection. In Journal of physics: Conference series (Vol. 1916, p. 012167).

Vidros, S., Kolias, C., Kambourakis, G., & Akoglu, L. (2017a). Automatic detection of online recruitment frauds: Characteristics, methods, and a public dataset. Future Internet, 9(1), 6.

Vo, M. T., Vo, A. H., Nguyen, T., Sharma, R., & Le, T. (2021). Dealing with the class imbalance problem in the detection of fake job descriptions. Computers, Materials & Continua, 68(1), 521–535.

Left Blank Intentionally

Public Perception about Climate Change and Its Impact in Different Seasons: A Micro-level Community Based Study in the South-Western Vulnerable Coastal Area, Bangladesh

Shamsunnahar Khanam[1*], Md. Golam Muktadir[1], Afnan Dilshad[1]

Abstract

The southwestern coastal zone of Bangladesh is the most vulnerable area to climate change because of its dynamic geomorphology, poverty and reliance of many livelihoods on climate sensitive sectors. The study investigates the key determinants of micro-level community perception of climate change in different seasons and dimensions as well as their source of predictions. In this study, we have applied both qualitative and quantitative methods. Data were collected from both primary and secondary sources. A field investigation was performed in two Upazilas of the Bagerhat district and primary data were collected through personal interviews of 65 respondents with the structured questionnaire checklist. And secondary data were collected from journals, reports, and historical meteorological data. Findings reveal that there is a sharp economical gap among the interviewed people and 96% of them perceive the increase of heat intensity and extension of summer. 65% of respondents find the shortening of the winter season and feel the 3^{o}C increase of minimum temperature that also revealed by the historical meteorological data analysis. The climate data also exhibits the 742.8 mm shortage of total rainfall during monsoon within a decade that is supported by 71% of people. Less than 25% of respondents find difficulties with groundwater but more than 50% complain the irrigation water dries up very soon. Almost all the respondents consider social media and TV news as well as their various senses are the source of their perception about local climate change. More than 90% of respondents perceive cyclones and salinity are their major hazards and also predict that storm surge along with sea level rise would add to this group. Policymakers should emphasize the outcomes of such study and design a zone wise adaptation plan that reflects public opinion, values, and demand.

Keywords: Climate change impacts, Micro-level community perception, Season change, Southwestern coastal zone

---

[1]Department of Environmental Science, Bangladesh University of Professionals, Dhaka, Bangladesh, Email: shamsun.nahar@bup.edu.bd

.

1. Introduction

There is very limited emphasis on the research of impact and perception of climate change at the community level however according to the UNEP's Global Environmental Outlook, climate change is the burning issue all over the world (Ravindranath and Sathaye, 2003). Climate change affects the natural ecosystems and socio-economic system of micro level communities, which are sometimes irreversible; result in changes at the macro level (Bhatti et al., 2006). Therefore, IPCC realize, recognize and focus on the impacts of climate change in different level and find that the communities of developing countries are more vulnerable and less adaptive to damages and other stresses related to climate change than developed countries (Maddison, 2007; Ravindranath and Sathaye, 2003). The effects of climate change are higher in magnitude in developing countries in terms of loss of life and relative effects on investment and the economy (Weigel, 1983). According to the Long-Term Climate Risk Index 2021 published by Germanwatch, Bangladesh is one of the most (7^{th}) affected countries by annual average natural calamities from 2000 to 2019 due to climate change (Kreft et al., 2019). Due to the global warming, the negative impact improves the ranking of Bangladesh in the context of increased and devastating floods in monsoon, increased drought due to reduced rainfall in winter, frequent cyclones and tropical storms, sea level rise, lower level of ground water, increased salinity, land and riverbank erosion (Chowdhury, 2011; Udwala, 2007).

The report published by MoEF in 2014 stated that impacts of climate change are not the same in every region of the country and south-western coastal districts are more vulnerable to climate change impacts because of the combined effect of gradual subsidence and sea level rise (Ahmed, 2010). The coastal zone of Bangladesh is the most vulnerable to climate change because of its geographic location, active delta and dynamic morphology, flat topography, high population density, high levels of poverty, and reliance of many livelihoods on climate sensitive sectors particularly-agriculture, fisheries and water resources (Karim and Mimura, 2008; Asada and Matsumoto, 2009; Smith and Leiserowitz, 2012). Climate change has a very adverse impact on the southwestern coastal and natural resources, agriculture, biodiversity, ecosystems, economy and widens the gap between rich and poor (Raghuvanshi et al., 2018). Eventually both rural and urban people of the coastal region of Bangladesh are more vulnerable to climate change because of their challenges of poverty, low infrastructural and technological development and high dependence on nature-based livelihood (Eriksson et al., 2009).

The perception of coastal vulnerable communities about the nature of climate change in different seasons, its intensity, impacts, and causes as well as the sources of their perception is crucial to explore (Mahmud and Uddin, 2017). Information on people perception is important for policy makers for a greater understanding of climate change impact and sustainable adaptive strategies, particularly in such vulnerable coastal communities (Hasan and Akhter, 2011).

People's realization of climate change form of human-environment interaction, functional ways of observing erratic changes, local causation behind changes in climate and its impacts. Thus, there is a growing need for a better understanding of the multi-faceted and complex linkages between climate change and public perception as well as the establishment of an international research community to address such issues.

The objective of the research is to listen, share and accommodate public knowledge, observation and perception about the effects of climate change in the southwestern coastal region. This study seeks to uncover the perception about climate change in various seasons like summer, winter and monsoon taking into consideration perception about agriculture, surface and ground water and the way people develop their perception regarding this.

2. Literature Reviews and Research Gap

The key researches, which were closely related to the focused study area and addressed the impacts of climate change, were reviewed for this study. Among them, Hussain & Ahmed (2020) examined the effectiveness of climate finance on the engagement of poor and marginal households in the southwestern coast of Bangladesh and found that climate finance recipient households engaged in more adaptation measures rather than others. An empirical study performed by Ali (2018) assessed that climate change was causing sea level rising and created salinity problem in the coastal areas of Bangladesh, which had caused damage to most of the agricultural production, fisheries and household and change of the occupation. He suggested that the developed countries that are responsible internally to increase climate change should support the affecting developing countries like Bangladesh to diminish climate change threat and also control over its phenomenon. Roy et al. (2020) explored the level and degree of practical implementation of the effectiveness of disaster risk reduction (DRR) processes in southwestern coastal Bangladesh. Their study addressed community perception and responsiveness as well as the role of both the local government and civil society and emphasized that all initiatives regarding disaster risk mitigation activities must involve and promote the local community. Rahman et al. (2013) found a substantial regional difference between coastal and other regions in terms

of continuous depletion and losses of resources. They revealed the significant regional disparity and lagging of the agriculture sector of this region by studying secondary data. Another study conducted by Islam et al. (2016) assessed the declination rice production in Bagerhat district due to the increased annual monsoon daily average and maximum temperature, seasonal total rainfall during 1989-2011 as well as the adverse effect of historical tropical cyclone "SIDR" held in 2007. Rahman and Kabir (2013) approached with the historical sequences in the landscape of saline tidal floodplain of Khulna and Shatkhira to understand the ecological settings and changes and indicated that the coastal embankments change the habitations, landscape and the settlement pattern which became vulnerable because of new hazards like a saline intrusion, tidal effects and flooding.

In the exiting literature, very limited information was found on the impact and perception of climate change at micro-level community focused. This is the first attempt at micro-level examining perceptions about climate change among the vulnerable southwestern coastal community in Bangladesh. Therefore, firstly, this study has filled up the dearth of the literature related to micro-level community in Bangladesh. Secondly, this study has examined perceptions from public opinion, which provide an understanding of the impact of regional, or global climate change in mangrove forest-based ecosystem along with the impact of climate change on micro-level community. And finally, this study finding will assist in further macro or regional level research for reviewing the relevant policies and institutional structures and frameworks and identifying the protection gaps.

3. Materials and Methods

Participatory Rural Appraisal (PRA) methods such as in-depth interviews along with previous literature reviews and historical meteorological data from open access were used to conduct the research, mentioned by Islam (2015).

The research was conducted through the sequential completion of respondent group selection, questionnaire design, and survey work and described both qualitative and quantitative ways. For quantitative analysis, primary data was collected from the local people in various villages (lowest rural geographic unit) of the selected Upazilas where household members who are adults (more than 18 years) and have literacy (ability to write a letter in any language) treated as the sampling unit.

A field investigation was performed in two phases in winter (20^{th}-25^{th} December 2019) and monsoon (10^{th}-16^{th} July 2020) and data had collected through personal interviews of 65 respondents with the structured questionnaire checklist related to

the indicators and perceptions in different seasons on the climate change (Mugambiwa and Dzomonda, 2018).

The respondents had selected through a multistage random sampling technique (Khan et al., 2018). The Climate Change Questionnaire (CCQ) was translated into the Bangla language by interviewers (Halady and Rao, 2010). All the questionnaire forms had been compiled, organized, checked carefully and irrelevant answers had been screened out. After analysing, the data extracted from primary sources were compared with the data from secondary sources. For secondary information, various types of open source data such as historical meteorological data, statistical data, and reports had been collected from various organizations.

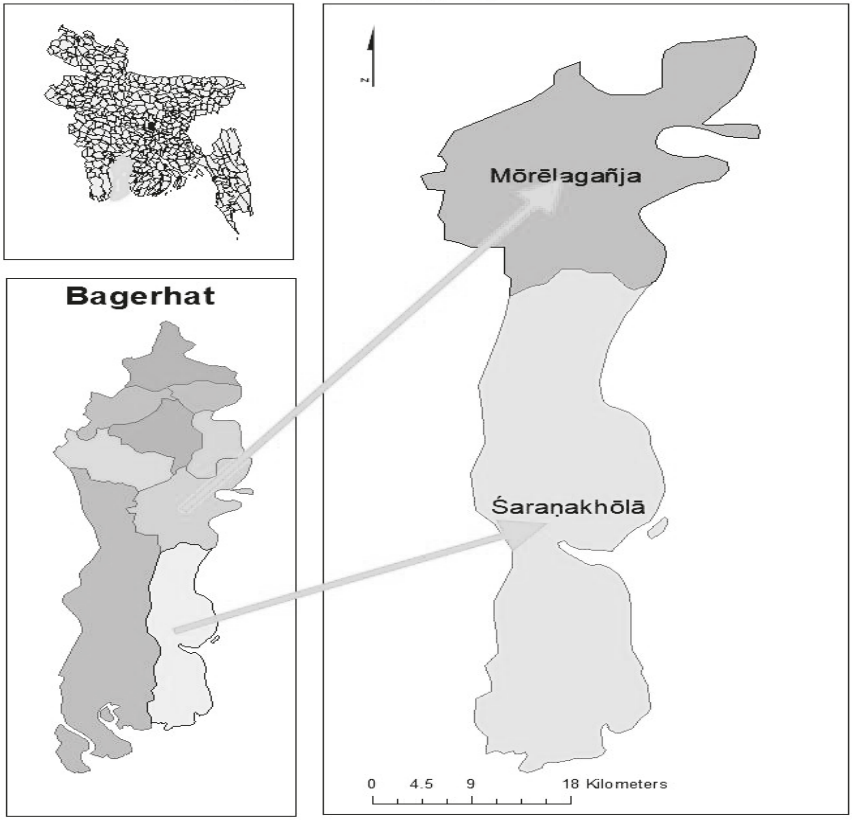


Figure 1: Location map of the study area (Morrelganj and Sarankhola)

.

But during the interview, no hints on climate change were provided to the respondents to avoid possible biasness as indicated by Mertz O, et al. (2011). As the discussion was made on their perceptions of climate change, they had made some changes on their original view of climate change and adaptive measures. Data were analysed through qualitative and interpretive ethnographic tradition with special reference to cases.

Two upazilas (**Figure 1**) were selected as representative based on their ecological, geographical, meteorological, demographical and socio-economic condition. Morrelganj is one of the densely populated (639/km^2) upazila of Bagerhat district with a small municipality whereas 80% area (594.58 km^2) of Sarankhola lies in the Sundarban reserve forest area (BBS, 2013).

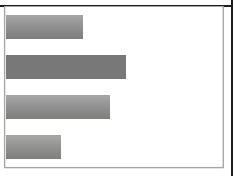
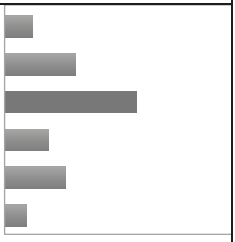
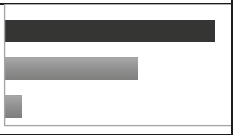
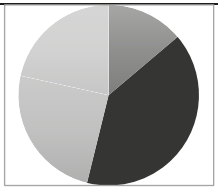
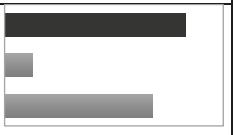
These two upazilas were selected purposively for the research because one has a semi-urban ecosystem and another has a rural mangrove forest ecosystem while both have a coastal riverine ecosystem. Administratively Morrelganj has four times (16) more unions than Sarankhola (4). The Pangunchi River cuts Morrelganj and its riverine area is 21.31 km^2. On the other hand, Sarankhola with a 7.78 km^2 riverine area is situated lengthwise west side Balaswar River, which separates the Khulna division from the Barisal division (BBS, 2013)

4. Result and Discussion

4.1 Qualitative Analysis/Descriptive Information of the Respondents

Patchen (2006) found that public perception, knowledge and attitude towards environmental issues varies on various personal characteristics such as gender, age, education, household size, occupation, family income and expenses which have a statistical effect on the probability of the respondents to perceive climate change (Khan et al., 2018). In Bangladesh, males are the head of the family and therefore most of the respondents involved in this sampling were males than females (Male=57%, Female=43%). Observing from the sampling frame in **Table 1**, the dominant sampling units (total 51 respondents) were above the age of 29 years, so that they could able to recall signs of patterns of climate changes in their communities.

Table -1: Demographic profile of survey respondents

		Total	Percentage	
Total number of respondents		N=65	100%	
Gender	Male	37	57%	
	Female	28	43%	
Age	19-28	14	21.54%	
	29-38	22	33.85%	
	39-48	19	29.23%	
	Over 49	10	15.39%	
Education	No formal education	5	7.69%	
	Class I-V	13	20%	
	Class VI-X	24	36.92%	
	Class XI-XII	8	12%	
	Graduation	11	16.92%	
	Post-graduation	4	6.15%	
Household	1-5 members	38	58.46%	
	6-10 members	24	36.92%	
	10+ members	3	4.61%	
Occupation	Agriculture	9	13.85%	
	Business	26	40%	
	Service	16	24.62%	
	Homemaker	14	21.54%	
Family Income (BDT/Month)	Less than 12,000	33	50.76%	
	Around 12,000	5	7.69%	
	More than 12,000	27	41.53%	
Family Expense (BDT/Month)	Less than 10,000	31	47.69%	
	Around 10,000	12	18.46%	
	More than 10,000	22	33.84%	

.

Most of the respondents had high school level education where some of them completed Junior School Certificate (JSC) and other passed Secondary School Certificate (SSC). In the context of the number of educational institutions and literacy level, Sarankhola is always lagging behind Morrelganj (BBS, 2013).

Now large family becomes very rare in the society of Bangladesh, which can be considered a great success of the GoB. Here only 4.61% of people belong to those families consisting of more than 10 persons. Due to the lack of agricultural scope and a good river-based communication system, it is found that a major portion of respondents earn their livelihood by doing business. There is an economical gap between rich and poor is evidently seen in the data frame where almost 50% of people earn less than 12000 BDT and expend less than 10000 BDT per month, which is the indirect impact of climate change on the socio-economic condition of a community. So, they are the potential resources for climate change study and should be explored to account for the impacts of climate change on coastal lives and livelihood.

4.2 Perception of Weather Changes Compared to Five to Ten Years Ago

4.2.1 Present Intensity of the Heat During Summer

Most of the people's perception of the intensity of heat is increased and 26% of men and 18% of women feel that the temperature during the summer season is high and very high respectively. Women are greatly concerned with climate change and consider that it is to be a near threat to men and act differently (McCright, 2010; Carlton and Jacobson, 2013). The historical data about the maximum temperature of the Bagerhat district **(Figure 3)** also trends high from 35oC in April 2011 to 37oC in April 2020 without some exceptionally high peak of 39oC in 2010 and 2014. This 2oC increment of maximum during summer aliens parallel with the public perception. People perceive the changes depending on some reasons like farmers are unable work long in their agricultural field due to extreme heat as well as 94% of people perceive the increment of heat intensity and 72% in heat waves and 36 respondents complain that the summer is extended occupying the days from pre-monsoon and spring **(Figure 2)**.

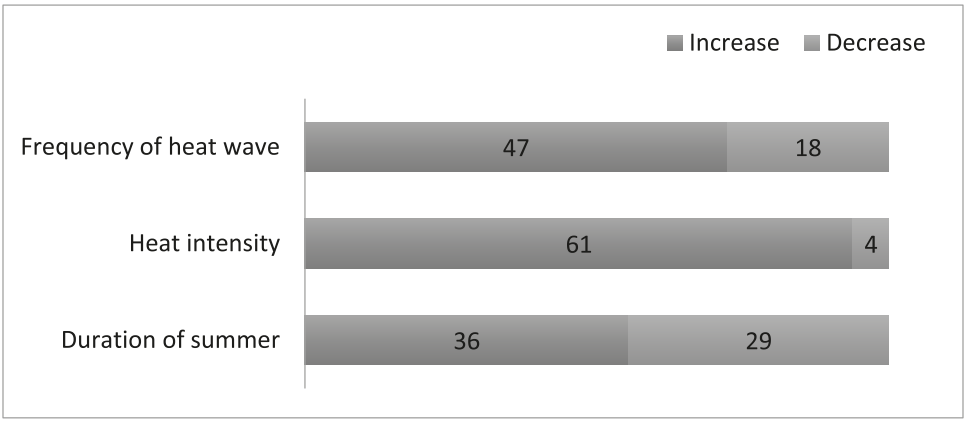


Figure 2: Perception of change in summer season depending on different phenomena

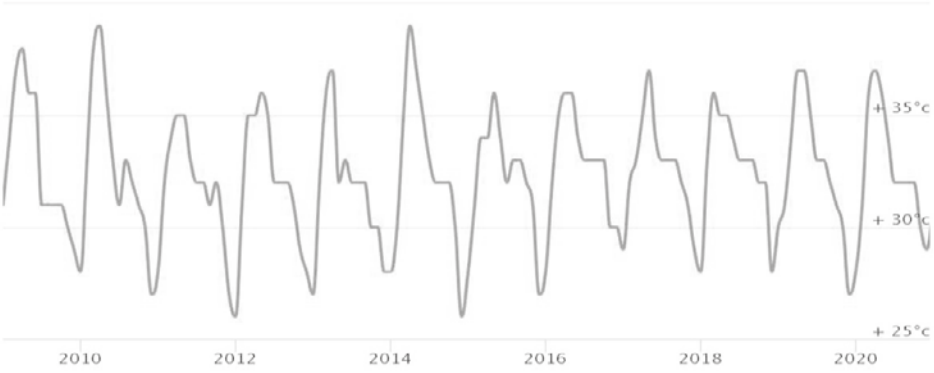


Figure 3: Graph of maximum temperature (^{0}C) from January 2010-December 2020 of Bagerhat district,Source:https://www.worldweatheronline.com/bagerhat-town-weather-averages/bd.aspx

4.2.2 Present Bitterness of Cold During Winter

The analysed meteorological data from figure 5 illustrates that the yearly minimum temperature of Bagerhat is gradually increased by 3^{o}C within this decade, which is shocking and alarming. In January 2010 the minimum temperature was 15^{o}C but it rises up to 18^{o}C, recorded in the same month of 2020. More than 65% of respondents find both the winter season and frequency of cold waves to be declined (**Figure 4**).

25% of men and 22% of women find that winter becomes warmer compared to five to ten years ago and 14% of men and 9% of women find it very low. 31 people out of 65 respondents do not notice the change of cold severity where 14% men and 8% women feel that the bitterness of winter becomes high. They may feel it due to the severe cold waves.

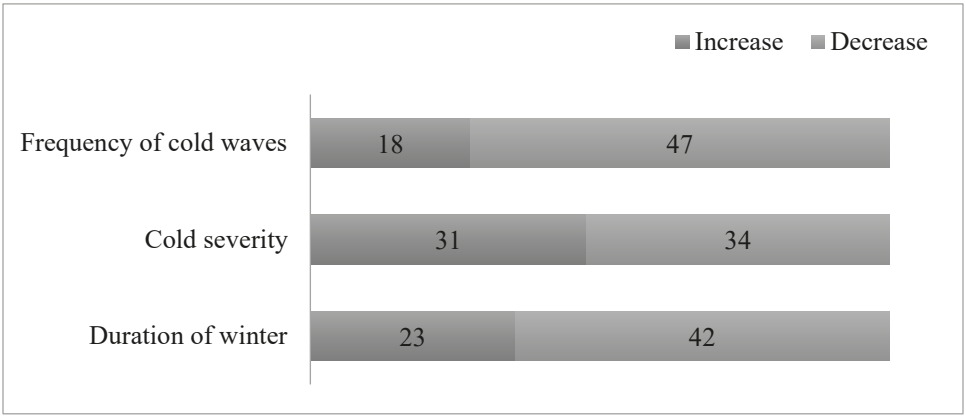


Figure 4: Perception of change in winter season depending on different phenomena

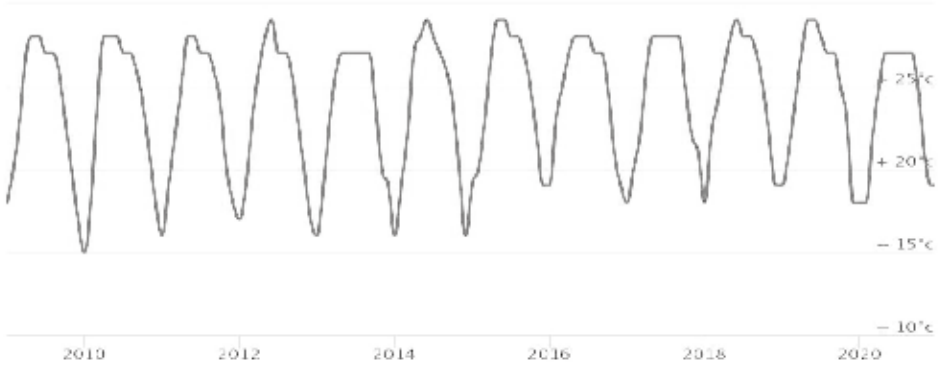


Figure 5: Graph of minimum temperature (^{0}C) from January 2010-December 2020 of Bagerhat district,Source:https://www.worldweatheronline.com/bagerhat-town-weather-averages/bd.aspx

4.2.3 Present Intensity of Rainfall During the Rainy Season

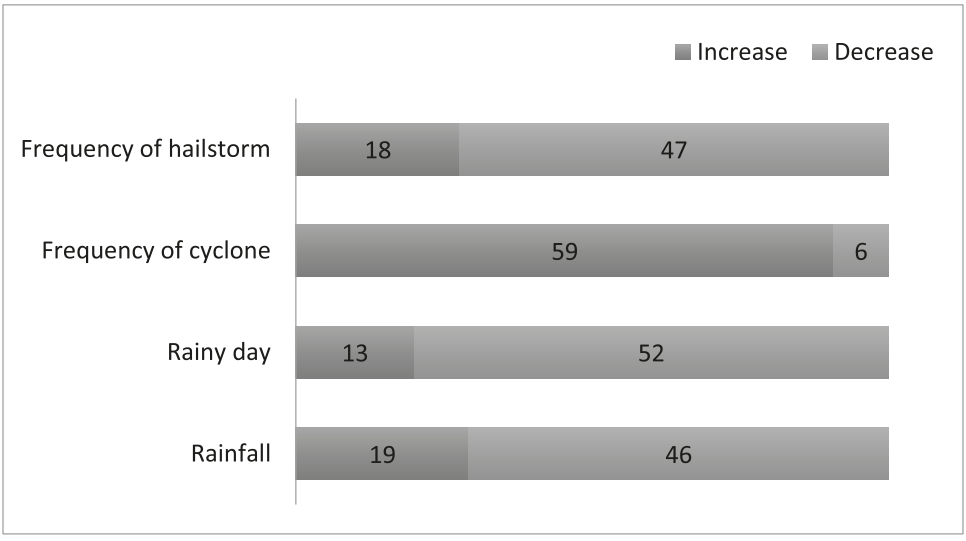


Figure 6: Perception of change in rainy season depending on different phenomena

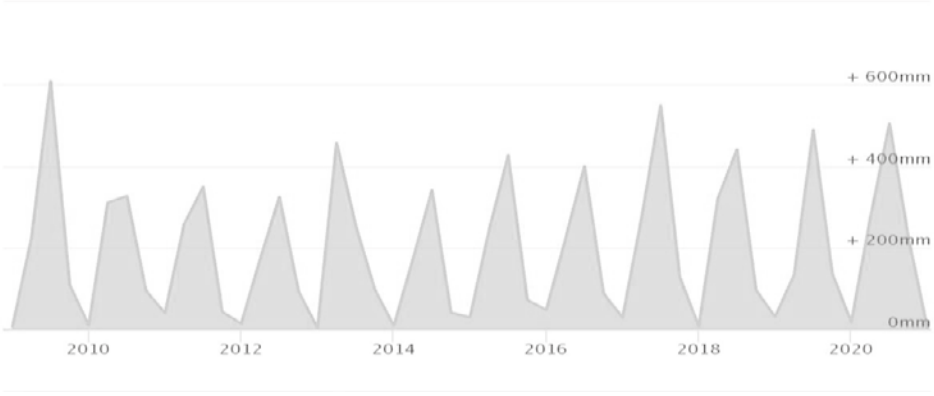


Figure 7: Graph of average rainfall (mm) from January 2010-December 2020 of Bagerhat district, Source: https://www.worldweatheronline.com/bagerhat-town-weather-averages/bd.aspx

Around 22% of male respondents perceive the intensity of rain is low which is lower than female respondents (25%). But few (8%) female think it remains normal where 20% of male support it. The average rainfall during monsoon in2009 (May-October) was 456.52 mm, which becomes 332.8 mm in 2019 (**Figure 7**) and 46 people support this decrement.

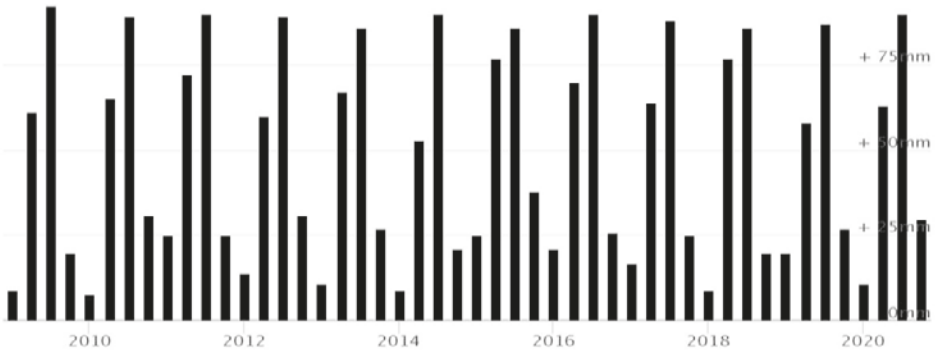


Figure 8: Graph of rainy days from January 2010-December 2020 of Bagerhat district, Source**:** https://www.worldweatheronline.com/bagerhat-town-weather-averages/bd.aspx

Table- 2: List of the Tropical Cyclone Hit Bangladesh in This Decade

Landfall Coast	Cyclone Name	Date	Month	Year	Category (SSHWS)	Wind Speed (km/h)
SW	Amphan	16-21	May	2020	5	260
	Bulbul	5-12	November	2019	3	195
	Fani	26-30	April	2019	5	280
	Titli	8-12	October	2018	3	195
SE	Mora	28-31	May	2017	1	150
	Roanu	19-23	May	2016		110
SW	Mahasen	16-21	May	2013		85
	Aila	26-27	May	2009	1	120
SE	Bijli	14-17	April	2009		95

Source: Joint Typhoon Warning Centre (JTWC)

Additionally, 52 people perceive the total rainy days during monsoon are also decreased (**Figure 8**). The historical rain data of the Bagerhat district (**Figure 8**) also exhibits that total rainy days were 159 days and 153 days respectively in 2009 and 2019. Although the total rainy days are decreased only 7 days, the amount of total rainfall is decreased by 742.8 mm during the rainy season. 47 people say that the frequency of hailstorms is decreased but maximum respondents (59 out of 65) agree that the frequency of cyclone in the southwestern coastal region has increased in high number because they are suffered almost every year by one or more category 3 to 5 tropical cyclone named Aila, Mahasen, Titli, Fani, Bulbul, and Amphan (**Table 2**). Their socio-economic damages are increased; however, they acknowledge that the Sundarban mangrove forest protected them as a shield and fatalities are minimized by early warning system.

4.2.4 Present Situation in Surrounding Water Resources

The water budget study by Zahid, et al. (2018) shows that the maximum depth to groundwater table in Bagerhat is not so changed under climate change condition because the river plays a significant role to adjust inflow and outflow. Around 58% of respondents find that there is no significant change to the water table in the tidal river but in ponds and canals during summer (**Figure 9**). This interconnection between the river and the aquifer provides drinking water to 75% of respondents where deep tube wells maintain good water pressure. A limited number of people (less than 25%) find water related difficulties in deep tube wells (**Figure 10**). The salinity level in southwestern coastal areas is increased with the coupled the effect of sea level rise (Khan et al., 2008) and shrimp farming. (Ali, 2006)

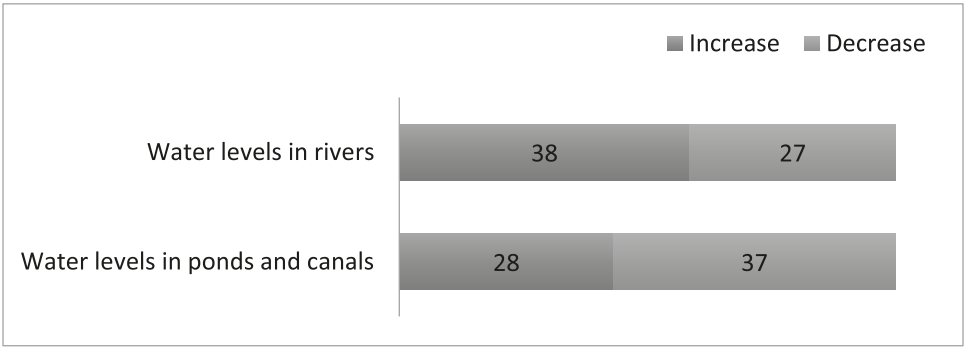


Figure 9: Perception change in surface water depending on different phenomena

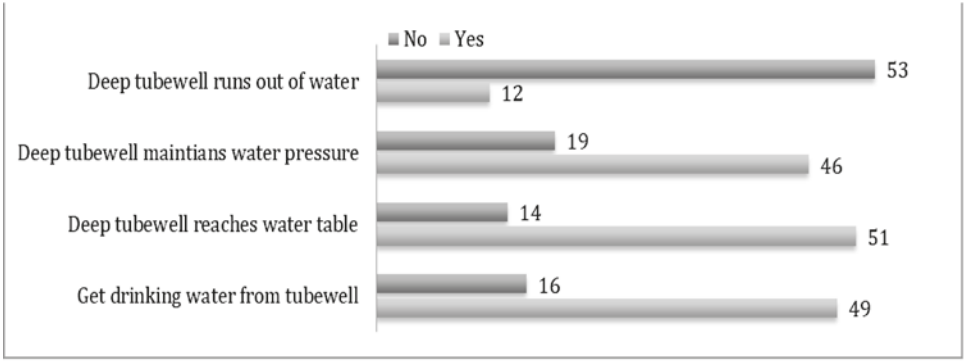


Figure 10: Perception change in groundwater depending on different phenomena

4.2.5. Present Situation in Agriculture

The change of agricultural characteristics plays one of the major roles regarding public perception about climate change. The respondents' perception regarding the impact of climate change in agriculture is identified through the effect of change in various phenomena related to cultivation and their return from farming (Khan et al., 2018). Response of sampled people revealed the majority of them (53) noted a significant decrease in the growth of the trees (**Figure 11**). More than 50% of people from total respondents admit that they can cultivate their land by rainwater but the irrigation water dries up very soon and filed become much drier. Moreover, cracks are formed in several agricultural fields and tender leaves of trees die in hot summer. However, a notable number of respondents (average 30) are not agreeing with that.

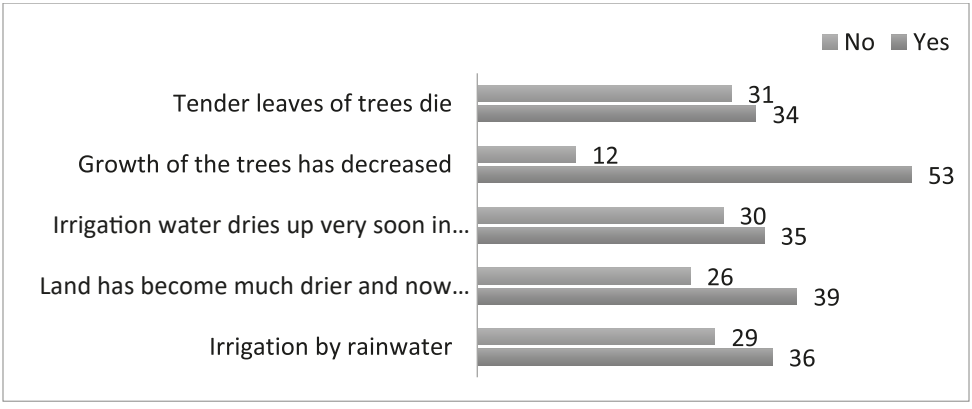


Figure 11: Perception change in agriculture depending on different phenomena

4.3 Perception about Present and Future Disasters

Present perception of the people from Morrelganj and Sarankhola indicates that cyclone and increased salinity in the surface and groundwater are the major hazards however all of these hazards are interrelated to each other results from climate change. They guess that storm surge would add this major group and drought; flood and sea level rise would remain the minor disasters as they are (**Figure 12**). Interestingly they think that water logging will no longer be their hazard, which is a subject to further detail research.

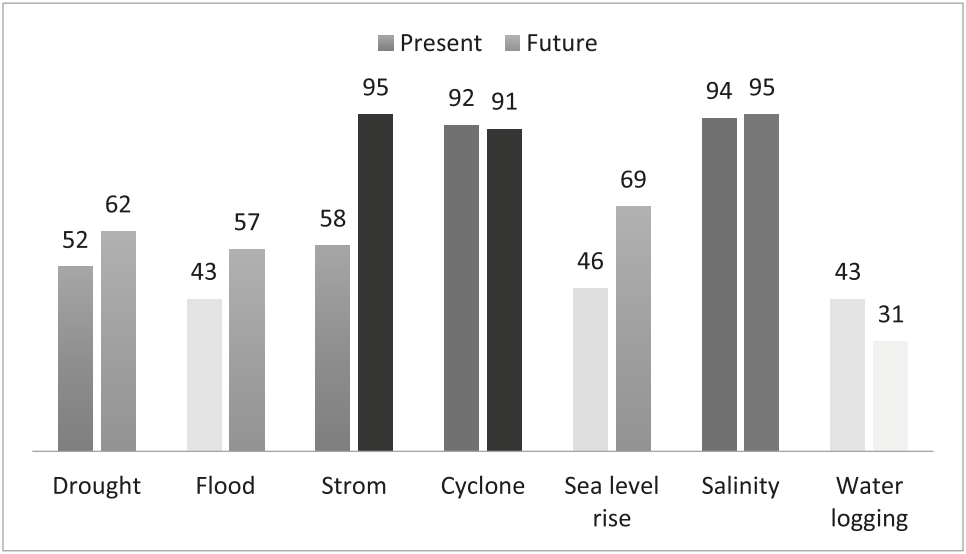


Figure 12: Local public perception about recent hazards and prediction about future hazards due to climate change

4.4 Source of Public Perception

All kinds of media play an important role in constructing people's concern about climate change and it also fluctuates with the amount of media coverage of a particular issue (Weigel, 1983; Mazrur and Lee, 1993). Most of the people (almost 60) get audio-visual evidence of changing global or regional weather pattern and their impact on the ecosystem by social media and TV news (**Figure 13**). Radio news is losing its popularity and newspaper works as a source for those people who have that much literacy level, patience and habit of reading.

Although different types of theoretical scale have been developed to measure people's psychological behaviours and environmental attitudes towards climate issues and changes (Maloney and Ward, 1973; Dunlap and Jones, 2003).

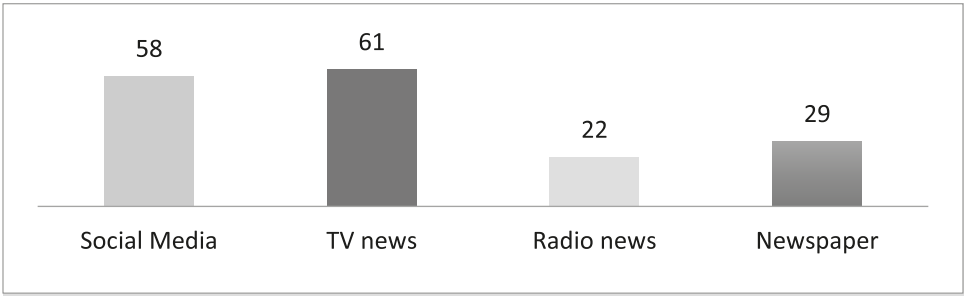


Figure 13: Various media sources of local public perception change on climate change

Feeling heat wave, tired and thirsty during summer while they are finding dewdrops presence with fog in winter are the main sensory source of their perception about climate change in their area (**Figure 14**). Some of them find winter arrives late comparing with previous years, with limited visibility of the sun due to dense fog. Even they cannot go outside of their house or feeling burning sensation and increased sweating because of extreme heat waves in summer.

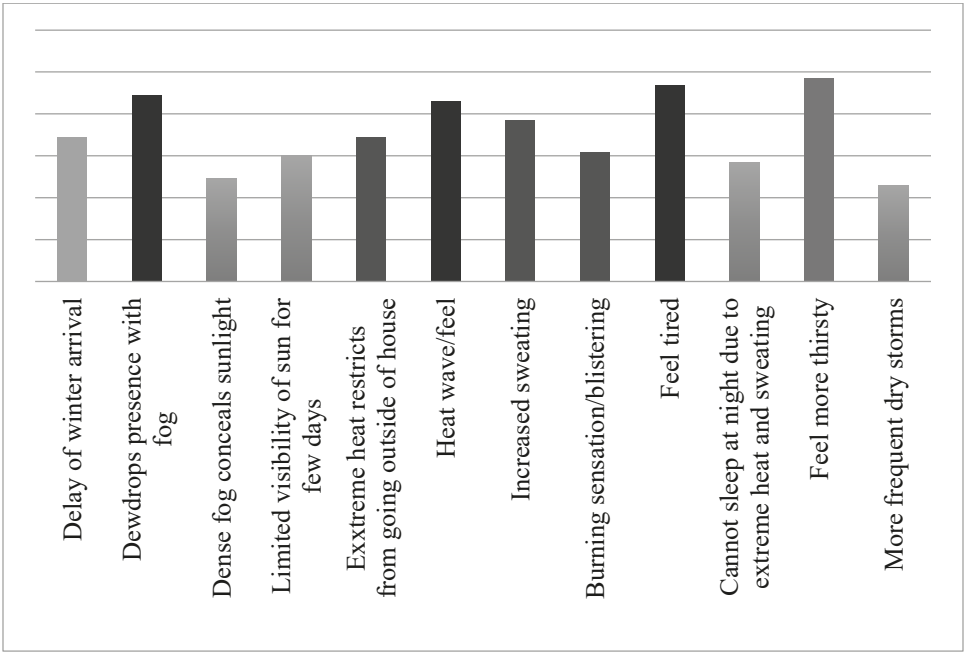


Figure 14: Various sensory sources of local public perception change on climate change

4.5 Public Perception about Climate Change

According to the Bagerhat district's meteorological statistics, the maximum summer temperature has increased by 2°C and the minimum winter temperature has decreased by 3°C over the past ten years. Rainfall and rainy days are both drastically declining during the monsoon season while rising during other seasons. Due to the blessings of the Balaswar River, despite the area being close to a coastal zone, where salinity in both surface and groundwater is rising and water quality is declining, there is no adverse effect on its hydromorphology for climatic change. People who live in this southwestern coastal region are extremely exposed to many kinds of natural disasters and risks, and almost every year they experience one or more category 3 to 5 tropical cyclones. They believe that because of the existing issue of increased salinity in their soil and water, sea level rise will worsen their misery. Moreover, due to the loss of irrigation water from ponds and canals, agricultural land is rapidly drying up and crop growth is slowing down. In addition to sensory information, social media, and television news, these underprivileged people also observe longer summers and shorter winters, an increase in cyclone activity, and a decrease in monsoon rain. The community is exposed to the effects of climate change, and its capacity for resilience is quite low.

5. Limitations, Conclusion and Recommendation

Our study has several strengths but the outcomes should not be generalized because of the small sample size and shorter period and area of study. One limitation of this empirical research is that it was based on the expressed information and opinion of the respondents, which may not be free from individual biases and prejudices. Further large-scale research studies on this topic need to be undertaken using a large sized sample across the southwestern coastal region of Bangladesh.

There is a consensus of perceived climate change variability among studies conducted in coastal areas, which found that respondents perceived a general increase in temperature, the intensity of heat, cyclones, duration of summer and a decrease in rainfall, winter season and agricultural production. Respondents' perceptions are consistent with meteorological data. They perceive that all aspects of their entire life are disrupted by hazards and negative impacts of climate change. They also think that the Sundarban mangrove forest, which is a blessing for them works as a shield protecting them and the early warning system by the government through social media and TV news minimize the fatalities. However, the poor infrastructure and resilience system increases the economic damages.

This micro-level study finding will assist in further macro or regional level research for reviewing the relevant policies and institutional structures and frameworks and identifying the protection gaps. Policymakers should emphasis on the outcomes of such study and design zone wise adaptation plan that reflects public opinion, values, and demand. Therefore, it is the utmost priority to be aware, resilient and protect these disregarded and afflicted vulnerable people of the southwestern coastal region from the climate induced hazards and problems.

References

Aftab K., Ali, S. S., Shah, S. A., & Khan, S. U. (2019). Determinants of farmers' perception about climate change in Khyber Pakhtunkhwa-Pakistan. Pure and Applied Biology. Vol. 8, Issue 4, pp2159-2168. http://dx.doi.org/10.19045/bspab.2019.80161

Ahmed, A. U. (Ed.). (2010). Reducing Vulnerability to Climate Change: The Pioneering Example of Community Based Adaptation in Bangladesh. Centre for Global Change (CGC) and CARE Bangladesh, Dhaka, p. 156.

Ali, A. M. S. (2006). Rice to shrimp: Land use/land cover changes and soil degradation in Southwestern Bangladesh. Land Use Policy, 23(4), 421-435. https://doi.org/10.1016/j.landusepol.2005.02.001

Annual Tropical Cyclone Reports (2020), Naval Oceanography Portal, U.S. Naval Meteorology and Oceanography Command https://www.metoc.navy.mil/jtwc/jtwc.html?cyclone

Asada H, Matsumoto J (2009) Effects of rainfall variation on rice production in the Ganges-Brahmaputra Basin. Clim Res 38: pp 249–260

Bhatti, J.S. & Apps, M.J. & Lal, R. (2006). Interaction between Climate Change and Greenhouse Gas Emissions from Managed Ecosystems in Canada. Taylor & Francis Group, LLC

Carlton SJ, Jacobson SK (2013) Climate change and coastal environmental risk perceptions in Florida. J Environ Manag 130:32–39. https://doi.org/10.1016/j.jenvman.2013.08.038

Chowdhury, G. A. (Ed.). (2011). Climate Change Challenges in Bangladesh. Centre for Environmental and Geographic Information Services, pp. 145.

District Statistics 2011: Bagerhat, (December, 2013), Bangladesh Bureau of Statistics (BBS)

Dunlap, R. & Jones, R. 2003. Environmental attitudes and values. Encyclopedia of psychological assessment, 1, pp 364-369.

Eriksson M, Xu J, Shrestha AB, Vaidya RA, Santosh N, Sandström K (2009) The changing Himalayas: impact of climate change on water resources and livelihoods in the greater Himalayas: International centre for integrated mountain development (ICIMOD).

Halady, I. R., & Rao, P. H., (2010) Does awareness to climate change lead to behavioural change? International Journal of Climate Change Strategies and Management, 2(1), pp. 6-22.

Hasan, Z. & Akhter, S. 2011. Determinants of public awareness and attitudes on climate change in urban Bangladesh: Dhaka as a case. European Journal of Social Sciences, 21, pp 154-162.

Hussain, Firdaus & Ahmad, Mokbul. (2020). Effect of climate finance on adaptation in the southwestern coastal region of Bangladesh. Development in Practice. 30. pp 1-18. 10.1080/09614524.2020.1779663.

Islam, Mohammad. (2015). Community Based Adaptation to Climate Change in the Exposed Coastal Areas of Bangladesh.

Islam, M M, Faruk, M. A. & Baten M A. (2016). Impacts of Some Climatic Variables on Productivity of Boro rice in South-Western Coastal Regions of Bangladesh Journal of Environmental Science and Natural Resources 9(1): pp 95-98 DOI:10.3329/jesnr.v9i1.30299

Karim MF, Mimura N (2008) Impacts of climate change and sea-level rise on cyclonic storm surge floods in Bangladesh. Glob Environ Change 18:pp 490–500. https://doi.org/10.1016/j.gloenvcha.2008.05.002

Khan, A., Mojumder, S. K., Kovats, S., & Vineis, P. (2008). Saline contamination of drinking water in Bangladesh. The Lancet, 371(9610), P385. https://doi.org/10.1016/S0140-6736(08)60197-X

Kreft, Sönke; David Eckstein, David; Melchior, Inga (December 2019). Global Climate Risk Index 2020 (PDF). Bonn: Germanwatch e.V. ISBN 978-3-943704-77-8. Archived (PDF) from the original on 25 December 2019. Retrieved 9 December2020.

Maddison D (2007) The perception of and adaptation to climate change in Africa: The World Bank.

Mahmud, S M & Uddin, Md. Borhan. (2017). Local Interpretations of Climate Change: A Study in Selected Coastal Areas of Bangladesh.

Maloney, M. & Ward, M. 1973. Ecology: Let's hear from the people: an objective scale for the measurement of ecological attitudes and knowledge. American Psychologist, 28, pp 583-586.

Mazrur, A. & Lee, J. 1993. Sounding the global alarm: environmental issues in the US national news. Social Studies of Science, 23, pp 681-720.

McCright AM (2010) The effects of gender on climate change knowledge and concern in the American pub- lic. Popul Environ 32: pp 66–87. https://doi.org/10.1007/s11111-010-0113-1

Mertz O, Mbow C, Reenberg A, Genesio L, Lambin EF, et al. (2011) Adaptation strategies and climate vulnerability in the Sudano-Sahelian region of West Africa. Atmosph Sci Let 12(1): pp 104-108.

MOEF 2014. BANGLADESH CLIMATE CHANGE STRATEGY AND ACTION PLAN 2014. Ministry of Environment and Forest, Government of Bangladesh.

Mugambiwa SS, Dzomonda O (2018) Climate change and vulnerability discourse by students at a South African university. J Disaster Risk Stud 10: pp 1–6. https://doi.org/10.4102/jamba.v10i1.476

Nawazeesh Muhammad Ali (2018). Impact of climate change and natural catastrophe on the occupational changes in the coastal areas of Bangladesh: an empirical study. Environmental Economics (open-access), 9(1), pp 22-37. doi:10.21511/ee.09(1).2018.02

N.H. Ravindranath, Jayant A. Sathaye. (2003). Climate Change and Developing Countries, Kluwer Academic Publishers, eBook ISBN: 0-306-47980-X, page 63-64

Patchen, M. (2006). Public attitudes and behaviour about climate change: what shape them and how to influence them. East Lafayette, Indiana, University of Purdue.

Raghuvanshi, Rupan & Ansari, Mohammad Aslam. (2018). Farmers Perception about Climate Change and their Knowledge of Adaptation Strategies.16. pp 1-7.

Rahman, Mashrur & Zaman, Md Kawser Bin. (2013). Agricultural Vulnerability in the Context of Coastal Region of Bangladesh. Bangladesh Research Publications Journal. 9.pp 64-68.

Rahman, Tahmina & Kabir, Khondaker. (2013). Effects of Embankments on the Habitation of the Southwestern Coastal Region of Bangladesh. AJEDM. 24.pp 65-80. 10.3850/S179392402013001774.

Roy, Tuhin & Nasreen, Mahbuba & Khan, Abdullah & Karim, Md. Rezaul. (2020). Disaster Risk Mitigation Strategies in the Southwestern Coastal Region of Bangladesh. 10.1007/978-3-030-23796-7_5.

Salauddin M, Ashikuzzaman M (2012) Nature and extent of population displacement due to climate change triggered disasters in south-western coastal region of Bangladesh. Int J Clim Change Strateg Manag 4: pp 54–65. https://doi.org/10.1108/17568691211200218

Smith N, Leiserowitz A (2012) The rise of global warming scepticism: exploring affective image associations in the United States over time. Risk Anal 32: pp 1021–1032. https://doi.org/10.111 1/j.1539-6924.2012.01801.x

Titumir, R.A.M. and Basak, J.K. (2012). Effects of Climate Change on Crop Production and Climate Adaptive Techniques for Agriculture in Bangladesh. Social Science Review, 29(1), pp 215-232.

Udwala, T. 2007. Role of media on climate change stressed. The financial express, May 31.

UNDP Human Development Report (2008) Impacts of Climate Change on Human Development, 2008. Available at http://hdr.undp.org/en/content/impacts-climate-change-human-development [Accessed on 13 February, 2018]

Weigel, R. 1983. Environmental attitudes and the prediction of behaviour. Environmental psychology: Directions and perspectives, pp 257-287.

World Bank. (2000). Bangladesh: Climate Change and Sustainable Development. Report No. 21104-BD, Rural Development Unit, South Asia Region, The World Bank (WB), Dhaka, pp. 95.

https://www.worldweatheronline.com/bagerhat-town-weather-averages/bd.aspx

Zahid, Anwar & Hossain, A. & Mridha, Gautam & Hasan, M. & Islam, Kamrul & Abbassi, M. (2018). Model Impact of Climate Change on the Groundwater Flow and Salinity Encroachment in the Coastal Areas of Bangladesh. 10.1007/978-981-10-3889-1_32.

Impacts of Tannery Effluents on the Environmental Quality of Hazaribagh Area of Bangladesh and its Possible Remediation Measures: A Review

Faria Jahan[1], Md. Arifur Rahman Bhuiyan[1*], Israt Jahan[1], Md. Golam Muktadir[1], Fatema Shahinur Jahan[2]

Abstract

Tannery effluents are containing harmful inorganic chemical substances such as chlorides, sulphides, chromium, lead, mercury, nitrogenous compounds and tannins as well as trace organic chemicals. The increased use of synthetic chemicals like dyes and finishing agents are mainly responsible for these pollutants. In the last few decades, pollution from tanneries at the leading industrial site in Dhaka discharged into the Buriganga River. These substances are commonly toxic and persistent, posing health and environmental risks. In this article, we discussed the deterioration of environmental quality with tannery effluents. The health impacts, water quality, damaging ecology, and threatening people's livelihoods are directly and indirectly related to exposure to hazardous wastes, including carcinogenic effects, reproductive system damage, respiratory system effects and central nervous system effects. The goal of this review was to evaluate the environmental quality in the Hazaribagh tannery industrial area. This study discusses many remediation approaches that help improve environmental quality. These findings recommend monitoring and cleaning up tannery effluent simultaneously. This review examined heavy metal contamination, exposure toxicity, research gaps, existing regulations, and long-term remediation methods in Hazaribagh to enhance environmental quality.

Keywords: Buriganga River, Environmental Quality, Heavy Metals, Hazaribagh, Tannery Effluents.

1. Introduction

"Environmental Quality" refers to environmental properties and characteristics that affect humans and other species. It measures the condition of an environment based on the needs of one or more species or humans. Tannery effluents rank first among all industrial wastes in terms of pollution (Haque et al., 2019). In Bangladesh, the tanning industry is considered one of the oldest industries. About 95% of the total tanneries of the country are situated in Hazaribagh amidst a densely populated residential zone. In 2003, the government had taken the

[1] Department of Environmental Science, Faculty of Science and Technology, Bangladesh University of Professionals, Dhaka, Bangladesh, Email: fariajahan499@gmail.com, arifur.rahman@bup.edu.bd*, isratjahan@bup.edu.bd, golam.muktadir@bup.edu.bd

[2] TÜV Rheinland Bangladesh Pvt. Ltd., Email: fatema.jahan@bd.tuv.com

initiative to shift to Savar but still, there have been some tanneries and previous activities left some long-term impacts.

The capital city and the river Buriganga are badly damaged by these hazardous tannery effluents (Mondol et al., 2017). A persistent pollution issue threatens the Buriganga River. Pollutant sources in the Buriganga include mill and industrial waste, domestic garbage, clinical waste, sewage, dead animals, polymers, and oil. Eco-friendly solid and liquid wastes from tanneries are discharged directly into the Buriganga River and other related rivers without treatment (Zahid et al., 2006). The discharge of untreated tannery effluents into water bodies has long been a concern in Bangladesh's leather industry (Asaduzzaman et al., 2016). The soil becomes contaminated with hazardous heavy metals as a result of tanning processes, and crops accumulate heavy metals during cultivation. Solid waste from the tannery business, which contains a significant quantity of hazardous metal, is turned into protein concentrate and fed to chickens (Islam et al., 2014).

Almost every tannery industry generates comprehensive use of chemicals during the conversion of animal hides to leather for tanning purposes (Chowdhury et al., 2015). Pesticides, tanning chemicals, and processed hides and skins handled in an inexperienced or dangerous manner may pose toxic hazards to human health (Cooman et al., 2003). Discharging effluents with high sodium, chloride, total dissolved solids, ammonia, nitrate, and sulphate concentrations revealed excessive salinity. Pollutants surpassed the threshold level due to an improper tanning process, causing toxicity in the receiving river water and rendering it unsuitable for fish and other aquatic species survival, as well as human and animal usage. Furthermore, when people utilized contaminated river water for irrigation, it depleted soil fertility and causes soil toxicity. Inorganic chemicals and heavy metals such as lead (Pb), chromium (Cr), cadmium (Cd), zinc (Zn), manganese (Mn), and arsenic (As) were found in tannery effluents. Tannery effluents were mainly utilized in agriculture, poultry feeding, and fishing operations, among other things. Harmful tannery effluents accumulate in human health due to these processes, causing a variety of illnesses.

This study is primarily a review work that was conducted by gathering information from previous studies. The environmental condition of the Hazaribagh industrial area over the last ten years has been critically reviewed here. The purpose of the study was to look into the effects of tannery waste disposal on the environment, specifically water, soil, vegetation, animals, agricultural production, and human health. This review study compared environmental quality to the standard index and identified appropriate and cost-effective solutions.

2. Data Collection

The review was based on an environmental quality assessment of a former tannery in Hazaribagh. In this review, investigations were conducted from 2014 to 2021 about health impacts, water quality, degrading ecological, and affecting people's livelihoods in Hazaribagh tannery industrial area. Relevant scientific publications identified major databases and original research papers on tannery effluents and remediation approaches. ScienceDirect, SpringerLink, PubMed, and Google Scholar were searched for relevant information. Search criteria included tannery effluent, environmental quality, heavy metals, Hazaribagh, and bioremediation approaches. Only 12 relevant papers were reviewed. Imitated information was available on the internet as very few research works were conducted. Updated information was found limited in number. The research methodology flow diagram has been shown in Figure 1.

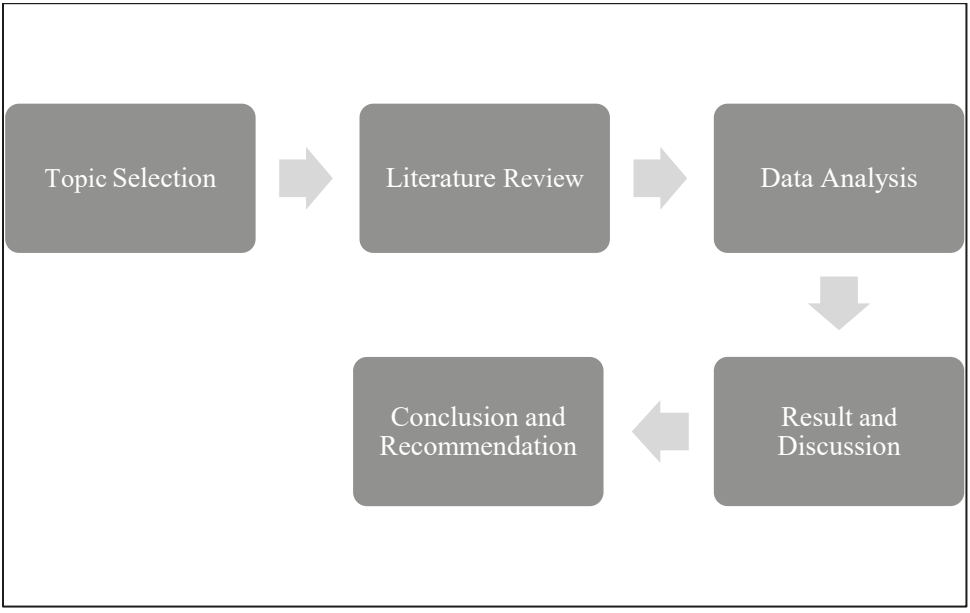


Figure 1: Research methodology flow diagram

3. Effects on Plants and Agriculture Production

Reducing tannery wastewater is a top priority for Dhaka City Corporation (DCC). Because tannery effluents are not treated, many hazardous elements end up in soils. Chromium accumulation in plants can reduce growth, photosynthesis, chlorophyll

content, enzyme activity and cause chloroplast and mitochondrial destruction (Guilizzoni, 1991). The levels of lead (Pb) and arsenic (As) in rice farming have already exceeded WHO and Egyptian standards. Salt stress slows metabolic processes or prevents the production of seed germination enzymes (Ashraf et al., 2002). The soil in the Hazaribagh area had high levels of Pb, Cd, and Zn. They were deposited in the soil, absorbed by nearby crops, and eventually entered the human body via the food chain. The carcinogenicity of Cr, Pb, and Cd had a variety of negative effects on human and animal health. As a result, authorities should act quickly to prevent heavy metal contamination (Rahaman et al., 2016). Heavy metal levels in Helencha (*Enhydra fluctuans*), Fern (*Nephrolepis exaltata*), and Rice (*Oryza sativa*) exceeded acceptable limits in Hazaribagh tannery industrial area (Sultana et al., 2015).

According to the streams of the literature, it was revealed that applying tannery effluents to soils with varying textures reduced rice yield, with the effect being more pronounced in light soils than in heavy soils. The effluent was also found to negatively impact growth performance. Metal content in soil experiment found heavy metal amounts. The condition of jute (*Corchorus capsularis*) and spinach (*Basella alba*) was observed by reference (Ali et al., 2015). Those jute (*Corchorus capsularis*) plant and spinach (*Basella alba*) was taken as a representative sample of plants in that region. Heavy metal concentration accumulations in vegetable leaf and shoot segments (Vine spinach) are presented in **Table 1** and contrasted, as far as possible, with FAO/WHO, India, and China for these components in fresh vegetables (Juel et al., 2016). Because of the high quantity of ammonia near the tannery in Hazaribagh, certain floras are becoming extinct (Hashem et al., 2014).

Table-1: Heavy Metal Accumulation in Plants Grown in the Study Area (mg/kg dry wt.).

Species	Parts of plants	Cd	Cu	Cr	Pb	Ni	Zn	References
Vine Spinach	Leaf	--	**12.15**	**3.95**	**10.3**	**1.8**	**69.75**	(Juel et al., 2016)
	Shoot	--	**10.75**	**8.1**	**13.7**	**1.55**	**40.35**	
Jute	Root	**4.6**	--	**176**	**15**	--	--	(Ali et al., 2015)
	Stems	**2.0**	--	**170**	**18.4**	--	--	
	Leaves	**0.212**	--	0.852	**0.348**	--	--	
Spinach	Root	**3.5**	--	**183**	**10**	--	--	(Ali et al., 2015)
	Stems	**5.2**	--	**190**	**12**	--	--	
	Leaves	**0.212**	--	0.261	**0.132**	--	--	
Vine Spinach	Edible	--	**16.91**	**64.88**	**13.86**	**2.36**	**231.98**	(Mizan et al., 2020)

Spinach *(Spinacia oleracea)*	Edible	**0.32**	--	**44.48**	**11.48**	--	--	(Islam et al., 2018)
Safe limit[a]		0.02	10	1.30	2	10	0.60	
Safe limit[b]		--	30	20	2.5	1.5	50	
Safe limit[c]		--	10	0.5	0.02	--	20	

Safe limit[a]: WHO (1996)

Safe limit[b]: Awasthi (2000)

Safe limit[c]: Chinese national food standards (2012)

Vegetables contain more zinc than the other metals studied in this study, acknowledging the bioavailability or mobility of zinc mentioned previous section. Cu and Ni concentrations, on the other hand, were more significant than safe limits by Awasthi (2000) & Chinese national food standards (2012). As a result, regular consumption grown in the research area is hazardous to anyone's health. As a result, eating these vegetables grown in the study area must have serious health consequences for humans. Vegetables are more susceptible to heavy metal and metalloid contamination due to their rapid proliferation and direct transfer of metals and metalloids to leafy portions (Chang et al., 2014).

4. Effects on Poultry Feed Manufacturing and Livestock Production Sector

The most frequent solid wastes produced in tannery sectors were skin trimmings, fleshing, keratin, chrome shaving, and buffing waste (Kanagaraj et al., 2006; Bari et al., 2015). Protein was the predominant constituent of that waste, which was converted into protein concentrate for use in chicken feed, fish feed, and the manufacturing of organic fertilizer. Skin-cut wastes (SCW) were tanned by slicing dry skins into thin slices (Bari et al., 2015).

Bari et al. (2015) estimated the target hazard quotient (THQ) and hazard index (HI). When THQ is less than one, the risk of non-carcinogenic side effects is considered minimal. There may be concerns regarding potential health problems linked with overexposure when it exceeds 1. The THQs might be averaged among contaminants to establish a hazard index (HI) to quantify the overall risk of adverse health effects from numerous metal exposures. The HI is made up of several THQs for various drugs or exposure techniques. The HI was employed as a screening criterion in this investigation to assess if heavy metals from contaminated poultry posed a serious threat to human health. In this work, the hazard indices for the toxic components Pb, Cd, and Cr were estimated. When the risk index reaches 1.0, it raises concerns about the potential for health problems (Khan et al., 2008).

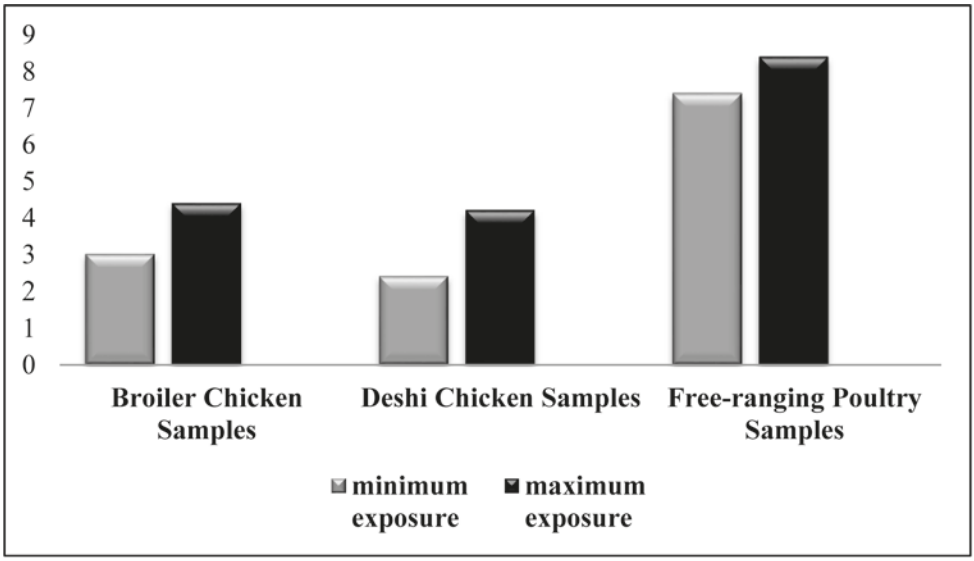


Figure 2: Target hazard quotient (THQ) for lead (Pb) Ingestion from Different Kinds of Chicken.

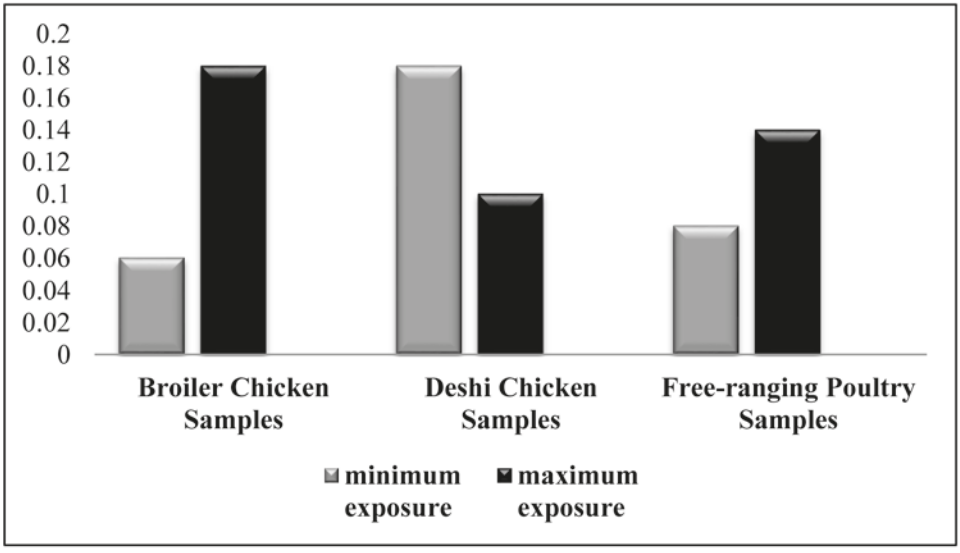


Figure 3: Target hazard quotient (THQ) for Cadmium (Cd) Ingestion from Different Kinds of Chicken.

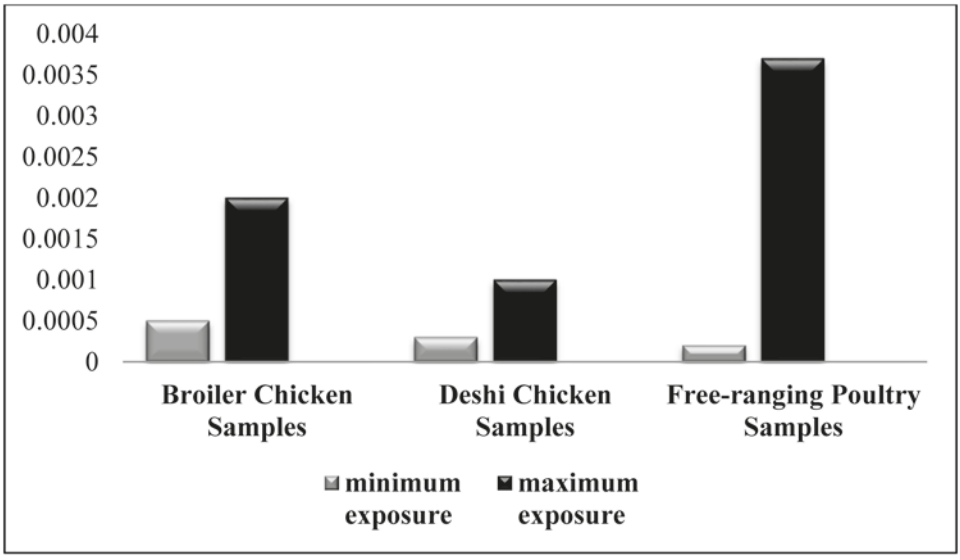


Figure 4: Target hazard quotient (THQ) for Chromium (Cr) ingestion from different kinds of chicken.

After analysing data (figure 2, 3, 4) from Bari et al. (2015), it was demonstrated that heavy metal contamination in feed enters into poultry and poses a potential risk to consumers through consumption of contaminated chicken meat.

Cadmium (Cd) and lead (Pb) levels from chicken foods had already exceeded WHO and Egyptian standard safe limits, indicating the health of the people in that area. Heavy metal contamination in feed penetrates birds, presenting a risk to humans who ingest contaminated chicken meat, according to analysis based on both studies (Islam et al., 2014).

Protein concentrates were made locally from tannery solid wastes, as well as other feedstocks was combined to provide a balanced diet. Two factors were considered. The first was the heavy metal content, which had been reduced due to dilution in the final feed. However, it made little difference where these protein concentrates were utilized to make poultry diets. The feed staffs, on the other hand, were combined with these proteins. Heavy metals may also be present in the feed staff (Hossain et al., 2007).

Tinni et al. (2014) conducted a survey on livestock production in the Hazaribagh tannery area and discovered that people involved in livestock production wanted to protect their animals from harmful chemical exposure but were unable to do so

due to excessive chemical deposition. A large number of livestock died as a result of the disease. Due to the discharge of hazardous tannery effluents, the animal farm in that survey region was found to be declining unnecessarily. Solid waste was dumped and discharged in public areas. Liquid waste is discharged without treatment in an unplanned manner. These harmful wastes were also consumed by livestock, which then become infected with various diseases. The liver, kidneys, and nervous system were all affected by the heavy metals found in tannery effluent. Chronic heavy metal exposure to the reproductive system is caused by steroid genic dysfunction, chromosomal defect and embryotoxicity (Verma et al., 2018).

5. Effects on Fisheries Production

Various studies indicate that tannery effluents are not treated before discharge into water bodies, harming aquatic resources and reducing fish productivity (Tinni et al., 2014). A study by Tinni et al. (2014) discovered that tannery effluent harms commercial fishing. Uncontrolled and untreated tannery waste discharge into open water bodies reduced fish productivity. In the Buriganga River, these wastes caused physio-chemical changes. Fish ingested tannery effluent components.

During the rainy season, polluted run-off river water entered the hatchery, killing the fish (Tinni et al., 2014). The presence and consequences of six heavy metals in tannery effluents discharged into the Buriganga River in Dhaka, Bangladesh, were reviewed by Asaduzzaman et al. (2016). Three fish species were sampled from the nearby river: climbing perch (*Anabas testudineus*), spotted snakehead (*Channa punctata*), and black tilapia (*Oreochromis mossambicus*). The experiment predicted that Cr and Pb were the most commonly accumulating elements in fish. Muscles of three fish species contained 2.70 mg/kg Cr. Those living closer to the study area ingested more Cr and Pb from eating local fish than those living farther away (Cd). Despite high levels of hazardous heavy metals in the river water, the fish species studied had concentrations far below the acceptable Food and Agriculture Organization/World Health Organization limits. They appeared to be safe to eat. Heavy metal concentrations were found to be within FAO and WHO guidelines. The results of the experiment show that fish have acceptable levels of heavy metals. This may change if higher authorities do not take preventative measures (Asaduzzaman et al., 2016). Polluted lead (Pb), chromium (Cr), cadmium (Cd), zinc (Zn), nickel (Ni), and manganese (Mn) were found in fish feed ingredients in Hazaribagh. The average chromium (Cr) concentration of all sample types was many times higher than the maximum allowed in the capital for tannery waste. Tannery waste can harm the liver and kidneys, and cause cancer if ingested

in fish and poultry (Akter et al., 2020). The chromium (Cr) concentration in Buriganga River fish was found to be higher than in fish from the local market. Heavy metal concentrations exceeded the WHO-approved threshold limit, indicating a serious health risk. They investigated *Heteropneustes fossilis* (stinging catfish) and *Channa punctata* (spotted snakehead) and found that metal content in fishes was higher than in sediment (Bashar et al., 2016).

Numerous studies have discovered that tannery effluents are not treated before discharge into bodies of water, resulting in harm to aquatic resources and decreased fish productivity, and one of those studies (Tinni et al., 2014) discovered that tannery effluents harm commercial fishing. Uncontrolled and untreated tannery waste discharge into open bodies of water reduced fish productivity. These wastes altered the physiochemistry of the Buriganga River.

The concentrations of heavy metals in fishes obtained in various studies were compared in **Table 2**.

Table-2: Targeted Heavy Metal Accumulation (mg/kg dry wt.) in Fisheries in the Study Area.

Species	Cr	Pb	Cd	Zn	Ni	Hg	Cu	As	References
Heteropneustes fossilis (Catfish)	**164.73**	**11.05**	**2.03**	**184.06**	--	--	--	--	(Bashar et al., 2016)
Channa punctata (Spotted snakehead)	**49.36**	**18.16**	0.717	**184.46**	--	--	--	--	(Bashar et al., 2016)
Channa punctata (Spotted snakehead)	**12.36**	**16.18**	**6.02**	--	--	**2.24**	--	3.39	(Islam et al., 2014)
Cirrhinus reba (Tatkini)	**12.89**	**15.76**	**6.03**	--	--	**1.67**	--	3.09	(Islam et al., 2014)
Oreochromis mossambicus (Mozambique tilapia)	**10.40**	**14.53**	**6.04**	--	--	**1.47**	--	3.10	(Islam et al., 2014)
Gudusia chapra (Chapila)	**6.31**	**10.92**	0.98	--	9.10	--	4.99	--	(Ahmad et al., 2010)
Glossogobius giuris (Baila)	**6.41**	**9.91**	0.87	--	9.7	--	5.03	--	(Ahmad et al., 2010)
Cirrhinus reba (Tatkeni)	**6.99**	**8.94**	0.87	--	9.60	--	4.33	--	(Ahmad et al., 2010)

Channa punctatus (Taki)	**5.65**	**9.10**	0.88	--	9.84	--	5.30	--	(Ahmad et al., 2010)
Mystus vittatus (Tengra)	**5.67**	**11.68**	1.13	--	9.41	--	4.31	--	(Ahmad et al., 2010)
Pseudeutropius atherinoides (Batashi)	**6.61**	**9.18**	1.04	--	9.15	--	4.85	--	(Ahmad et al., 2010)
Safe limits[d]	0.05	--	--	--	--	0.14	--	--	WHO (2004)
Safe limits[e]	--	1	2	0.05	--	1	--	--	Laws of Brunei (2001)

Fishes had higher concentrations of chromium (Cr), lead (Pb), and zinc (Zn) than the other metals studied in this study, confirming the bioavailability or mobility of chromium (Cr), lead (Pb), and zinc (Zn) indicated earlier in that article. However, quantities of chromium (Cr), lead (Pb), and zinc (Zn) were higher than FAO/WHO and Brunei guidelines. As a result, consuming these fish grown in the research area is harmful to human health. As a result, eating these fishes grown in the research location has severe health implications for humans.

6. Effects on Human Health

The main cause of health and safety violations is toxic chemical exposure in Bangladeshi tanneries (ILO, 2008). Skin and respiratory diseases are very common health problems among tannery workers as a result of exposure to hazardous chemicals (Azom et al., 2012). Tannery effluents can directly or indirectly affect human health and the food chain. Tannery effluents, which contain animal flesh, sulfuric acid, chromium, and lead, affect not only tannery workers but also the surrounding community (Biswas & Rahman, 2013). Tannery workers are thus potentially exposed to harmful agents and heavy metals, particularly Cr, which makes them vulnerable to health problems, particularly for those who have previously experienced respiratory tract and skin problems (Shahzad et al., 2006). Chromium hypersensitivity can lead to the production of complex antigens in tannery workers due to its ability to bind with skin proteins (Ali et al., 2015).

Hasan et al. (2016) found from the study that the common health problems (figure 6) among the tannery workers were Asthma (about 50% of workers), diarrhoea (71.7%), typhoid (43.5%), blood pressure (52.2%), gastrointestinal problems (71%) and eye problems (46.7%). In the case of occupational dermatitis among the tannery workers, the prevalence was found as scabies (73.9% of workers were

affected), nail discoloration (69.6%), urticaria (59.7%), miliaria and folliculitis (56.6%), contact dermatitis (39.13%).

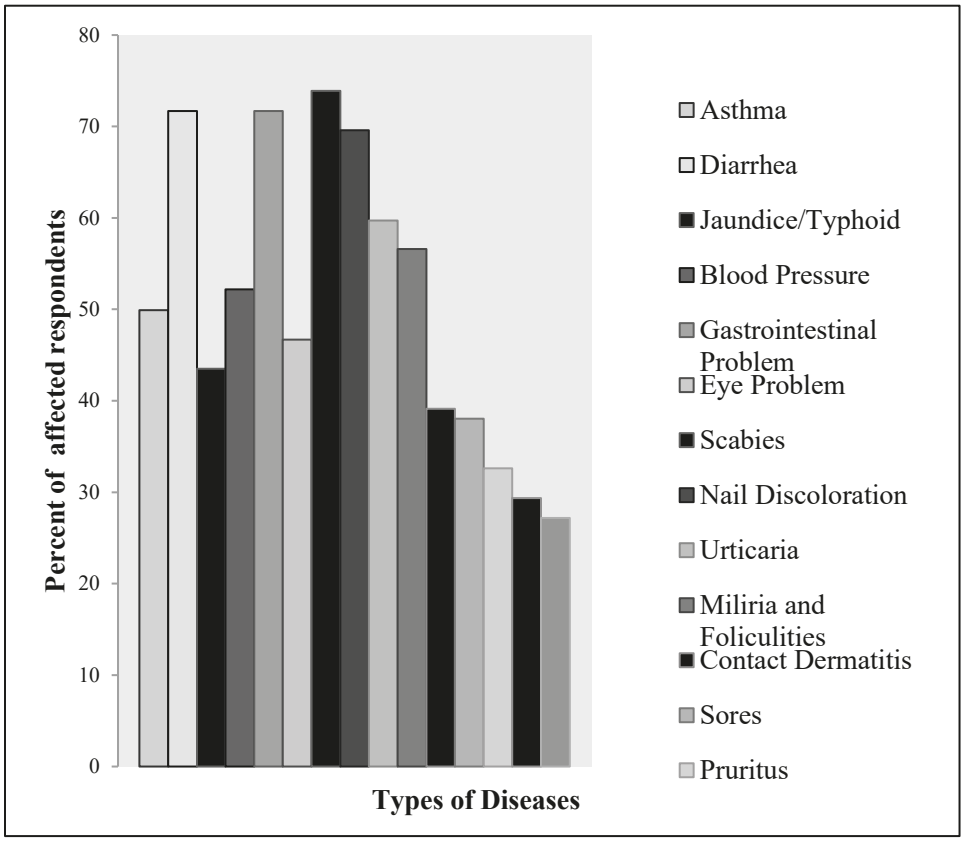


Figure 5: Percentage of affected people from suffering from various diseases.

The possibility of short-term health issues from chemicals like sulfuric acid and sodium sulphide has long existed. Several other chemicals, such as formaldehyde and azo colorants, have been linked to human cancer. They may swiftly invade the eye and cause long-term harm. During the deliming process, workers are exposed to high amounts of ammonia. With accessible acids in the environment, ammonia forms salts such as nitric acid (HNO_3), sulfuric acid (H_2SO_4), and hydrochloric acid (HCl). As a consequence, those who work in tanneries, whether wholly or partly, come into contact with ammonia or its hydroxide and experience a variety of problems. Thus, it causes sleepiness, nausea, and headaches (Hashem et al., 2014).

7. Effects on Reproductive Health

Das et al. (2015) surveyed that area and found some informative information about their reproduction health condition in the study area. From that survey, they found some important information about the reproductive health condition of the workers both male and female.

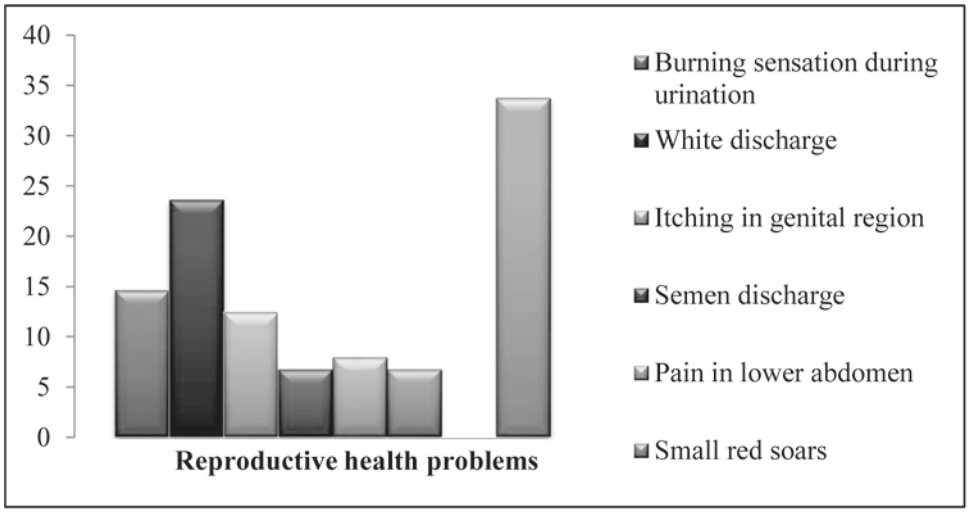


Figure 6: Reproductive health problems among female workers.

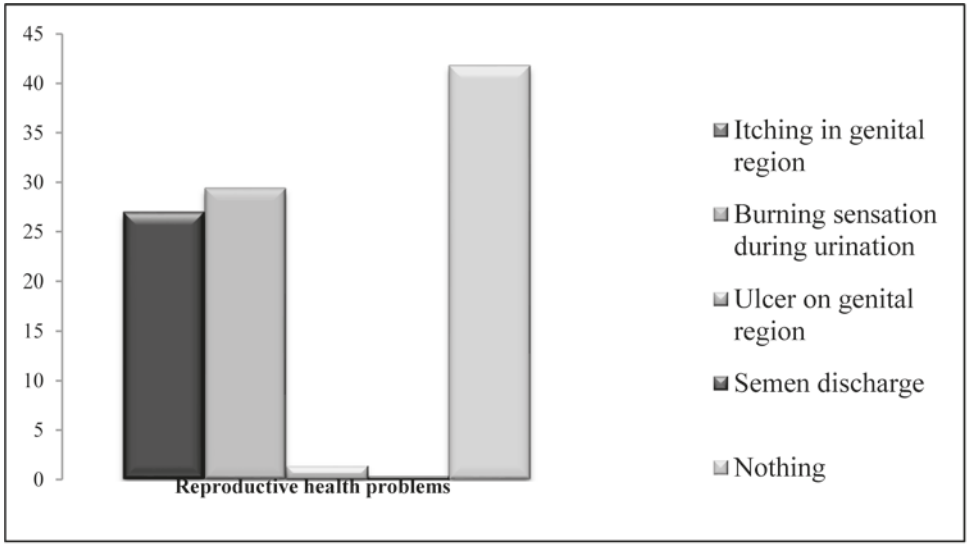


Figure 7: Reproductive health problems among male workers.

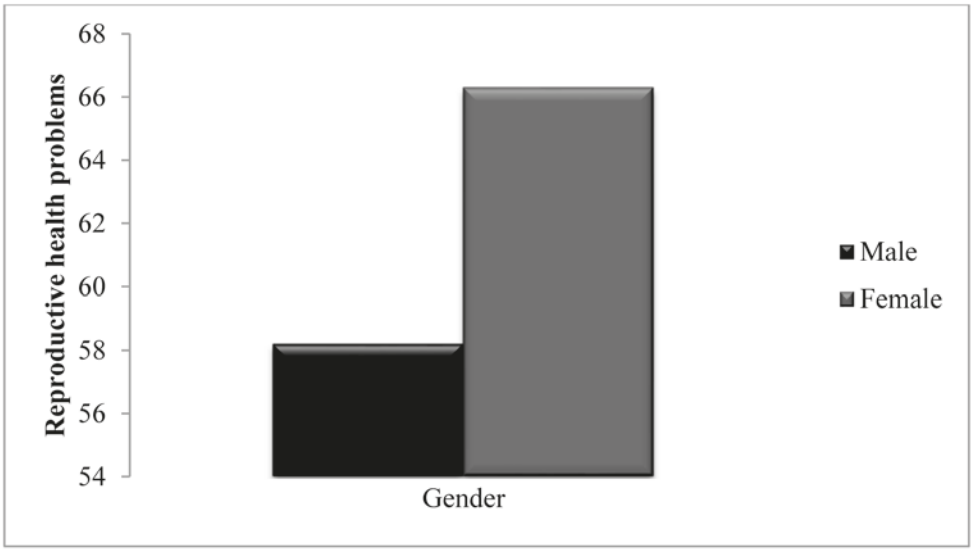


Figure 8: Percentage of reproductive health problems among male and female workers.

Burning sensations during urination, itching in the genital region, pus discharge from the urethra, ulcers on the penis, ulcers on the genital region, pain in the testis, warts on the genital region, and semen discharge were all reported by male workers (figure 7). It was clear that (figure 8) female reproductive health was more vulnerable than male reproductive health conditions.

According to the female participants, local pregnant women suffered from seizures during their pregnancies and had various reproductive issues. According to both male and female participants, environmental contamination caused by the incorrect disposal of a large volume of solid and liquid tannery wastes harms both tannery employees and residents of Hazaribagh (Haque et al., 2020).

8. Remediations of Tannery Effluents

The tannery sector consumes a lot of water and discharges a lot of tannery effluents that haven't been treated. Tannery effluent contains important concentrations of COD, BOD, TSS, TDS, and heavy metals. The traditional method of removing heavy metals have several drawbacks, including a high initial investment and operating cost, making it unsuitable for small-scale industries. One of the most important processes for removing heavy metals from tannery effluents is adsorption. Here this review work included two cheapest and most easily available

processes for removing heavy metals contain from tannery effluents such as using spent tea leaves and phytoremediation processes.

According to Alam et al. (2018), the surface water of the Hazaribagh tanning site has a large amount of Cr (939.81 mg/L). This study explained that tea waste is easily available in our country and it is a cheap material. The treatment included the use of waste black tea leaves as an absorbent material. According to Thakur & Parmar (2013), the percentage of removals of metals increases rapidly as the dose of the adsorbents increases due to the increased availability of exchangeable sites or surface area.

According to their research, tea waste can be used as a low-cost, locally and freely available, environmentally friendly, and efficient bio adsorbent for removing Cr from tannery wastewater. The maximum removal of Cr by spent tea leaves was 95.42 percent at a 14 g/L adsorbent dose and pH 10 in the experiment. Cr's maximum adsorption capacity on tea waste was discovered to be 10.64 mg/g (Alam et al., 2018).

Table-3: Techniques Used for Remediation.

Name	Benefits	Drawbacks	Examples
Phytoextraction	The biomass of the plant, which contains pollutant extract, can be reused as a resource; environmental friendliness.	The plant's biomass must be harvested and removed. Commonly, hyperaccumulator plants have shallow root systems, fewer roots and more biomass.	*Thlaspi caerulescens*, *Alyssum bertolonii*
Phyto stimulation	Chlorinated solvents have also been used to target locations for demonstrations. Exudates from the roots are used.	Plants can be harmed by high levels of pollution.	*hornwort*
Phytovolatilization	More harmful pollutants are transferred to less	Contaminants accumulated in flora, which then	*Astragalus bisulcatus*,

	toxic contaminants using this method. Toxic substances are released into the environment.	moved into fruits and other edible portions. In-plant tissues, the level of metabolites drops.	*Stanleya pinnata*
Phytostabilization	Low-cast, less disruptive than other soil remediation systems. Revegetation aids ecosystem restoration.	This process prevents the leaching and re-release of different contaminants, the contaminants must be kept in place. This requires careful maintenance.	*Eragrostis, Gladiolu, Alyssum*
Rhizodegradation	Ex-situ and in situ applications of this technology are feasible. The use of both territorial and aquatic plants is permissible here.	In this technology, continual pH adjustment is required to maintain the flow rate and concentration of influent, as well as a well-engineered system. Plant removal and harvesting are required.	*Pseudomonas*
Phytomining	This technology is useful for the uptake of precious metals, where conventional mining is not economically feasible	The disposal of reagents results in achieving precious metals is a threat to the environment.	*Alyssum bertolonii*

Source: Kamran et al., 2014

Cr contamination occurred as a result of tannery wastewater discharge into the Dhaleshwari River, and we identified potential native plants for Cr

phytoremediation. Additionally, samples of the root, stem, leaf, and fruit of four selected plants were collected from those sampling points (i.e., *Eichhornia crassipes*, *Xanthium strumarium L*., *Cynodon dactylon*, and *Croton bonplandianum Baill*.). *Xanthium strumarium L*. had the highest translocation factors (TF) and bioconcentration factors (BCF) for Cr of all the plant species studied. According to this study's findings, *Xanthium strumarium L.* is a more suitable native species for phytoremediation of Cr (Hasan et al., 2021).

Table-4: List of Plant Species from Bangladesh that are Capable to Remediate Potential Contaminants.

Plant species	Potential Contaminants	References
Fern (*Pteris vittata*)	Inorganics (As, F)	(Dey et al., 2016)
Xanthium strumarium	Cr	(Hasan et al., 2021)
Water pennyworth (*Hydrocotyle ranunculoides*)	Inorganics (As, F) and metals (Pb)	(Dey et al., 2016)
Amaranthus viridis L.	Pb	(Azam et al., 2014)
Vetiver grass (*Chrysopogon zizanioides*)	Pb, Cd	(Dey et al., 2016)
Azadirachta indica A. Juss.	Cu, Ni, Zn and Pb	(Azam et al., 2014)
Sunflower (*Helianthus annuus*)	Zn, Ni, Pb, Cu	(Dey et al., 2016)
Blechnum orientale L.	Pb, Ni	(Azam et al., 2014)
Marigold (*Tagetes patula*)		(Choudhury et al., 2015)
Indian mustard (*Brassica juncea*)	Cr, Cu, Pb and Zn	(Choudhury et al., 2015)
Water hyacinth (*Eichhornia crassipes*)	Pb, Cr, Cu, Hg, Zn, Cs, Sr, U	(Dey et al., 2016)
Commelina benghalensis L.	Cu, Pb, Cd, Zn	(Azam et al., 2014)
Brahmi (*Bacopa monnieri*)	Pb, Cd, Cr, Cu, Hg	(Dey et al., 2016)

Momordica charantia L.	Cu, Co, Cd, Fe	(Azam et al., 2014)
Pistia stratiotes L.	Cr, Co	(Azam et al., 2014)
Water lettuce (*Pistia stratiotes*)	Cu, Hg, Cr, Pb, Cd	(Dey et al., 2016)
Tridax procumbens L.	Hg	(Azam et al., 2014)

In Bangladesh, these plants can be found in abundance all around the country. The intrinsic property of these plants can be used to rehabilitate toxic material contaminated areas. The phytoremediation technique was shown to have a significantly higher decrease rate than the sedimentation technique (Chakrabarty et al., 2017).

9. Conclusion

The results obtained in this review study clearly depict that in Bangladesh, the uncontrolled and unplanned growth of tannery processing industries has contaminated land and water, raising concerns about public health. Tannery effluents are mainly used as pesticides in agricultural lands, poultry feeding elements, and fisheries sectors for these reasons public health facing vulnerable conditions. Discharged tannery effluents with heavy metals without any treatment cause many environmental and health effects. Consumption of heavy metals contaminated food results in long-term accumulation in the human body such as through bioaccumulation. After several years of exposure, it was discovered that several adverse effects on humans, including thalassemia, dermatitis, brain and kidney damage, and cancer, may be observed. Used tea leaves are a readily available that can be converted into a valuable product as an adsorbent for chromium (Cr) removal from tannery wastewater and phytoremediation which was less expensive. Based on this review study, further study should be conducted about environmental impact analysis over the former tannery area which is adjacent to the Buriganga River.

References

Ahmad, M., Islam, S., Rahman, S., Haque, M. and Islam, M., 2010. Heavy metals in water, sediment and some fishes of Buriganga River. Bangladesh. *International Journal of Environmental Research*, 4(2), pp. 321-332.

Akter, A., Mondol, M.N., Chamon, A.S. and Fiaz, S.M.A., 2020. Heavy Metals in poultry and fish feed ingredients in Bangladesh: A potential threat to our next generation. *International Journal of Engineering Applied Sciences and Technology*, 5(8), pp. 64-70.

Alam, M.N.E., Mia, M.A.S., Ahmad, F., & Rahman, M.M., 2018. Adsorption of chromium (Cr) from tannery wastewater using low-cost spent tea leaves adsorbent. *Applied Water Science*, 8(5).

Ali, H., Khan, E., & Sajad, M.A., 2013. Phytoremediation of heavy metals – Concepts and applications. *Chemosphere*, 91, pp. 869-881.

Ali, M.F., Naher, U.H.B., Chowdhury, A.M.S.U., Rahman, G.M.S. and Hasan, M.M., 2015. Investigation on physicochemical parameters of tannery effluent. *Universal Journal of Environmental Research and Technology*, 5, pp. 122-130.

Asaduzzaman, M., Hasan, I., Rajia, S.,Khan, N. and Kabir K.A., 2016. Impact of tannery effluents on the aquatic environment of the Buriganga River in Dhaka, Bangladesh. *Toxicology and Industrial Health*, 32(6), pp. 1106-13.

Ashraf, M.Y., Afaf, R., Qureshi, M.S., Sarwar, G. and Naqvi, M.H., 2002. Salinity induced changes in α-amylase and protease activities and associated metabolism in cotton varieties during germination and early seedling growth stages. *Acta Physiologiae Plantarum*, 24, pp. 37–44.

Awasthi S.K., 2000. Prevention of Food Adulteration Act No. 37 of 1954. Central and State Rules as Amended for 1999, third ed. Ashoka Law House, New Delhi. Azam, F.M.S, Haq, W.M., Rahman, S. and Jahan, S., 2014. Phytoremediation and Prospects of Cleaning up a Tannery Waste Contaminated Site in Dhaka, Bangladesh. Advances in Environmental Biology, 8(12), pp. 242-246.

Azom, M.R., Mahmud, K., Yahya, S.M. and Himon, S.B., 2012. Environmental Impact Assessment of tanneries: A case study of Hazaribagh in Bangladesh. *Int J Environ Sci Develop, 3*, pp. 152-156.

Bari, M.L., Simol, H.A, Khandoker, N., Begum, R., and Sultana, U.N., 2015. Potential human health risks of tannery waste-contaminated poultry feed.

Journal of Health and Pollution, *5*(9), pp. 68-77.

Bashar, K., Afroze, J.S., Ali, M.F., Rahaman, A. and Hosen, M.R., 2016. Relative study of heavy metal contamination in fishes of Buriganga River due to untreated tannery effluent and in fishes from another source in Bangladesh. *International Journal of Current Research, 8*(10), pp. 40395-40403.

Biswas, S. and Rahman, T., 2013. The effect of working place on worker's health in a tannery in Bangladesh. *Adv Anthropol 3*, pp. 46–53.

Chakrabarty, T., Afrin, R., Mia, M.Y. and Hossen, M.Z., 2017. Phytoremediation of Chromium and some chemical parameters from tannery effluent by using Water Hyacinth (*Eichhornia craassipes*). *Res. Agric. Livest. Fish*, *4*(3), pp. 151-156.

Chang, C.Y., Yu, H.Y., Chen, J.J., Li, F.B., Zhang, H.H. and Liu, C.P., 2014. Accumulation of heavy metals in leaf vegetables from agricultural soils and associated potential health risks in the Pearl River Delta, South China. *Environ Monit Assess*, *186*, pp. 1547–1560.

Chinese National Food Standards. 2012. 12th Five-year Plan on National Food Safety Standards.

Choudhury, M.R., Islam, M.S., Ahmed, Z.U. and Nayar, F., 2015. Phytoremediation of heavy metal contaminated Buriganga riverbed sediment by Indian mustard and marigold plants. *Environmental Progress & Sustainable Energy, 35*(1), pp. 117–124.

Chowdhury, M., Mostafa, M.G., Biswas, T, Mandal, A. and Saha, A., 2015. Characterization of the effluents from leather processing industries. *Environ. Process*, *2*, pp. 173-187.

Cooman, K., Gajardo, M., Nieto, J., Bornhardt, C. and Vidal, G., 2003. Tannery wastewater characterization and toxicity effects on Daphnia spp. *Environmental Toxicology, 18*(1), pp. 45–51.

Das, A.C., Sultana, R. and Sultana, S., 2015. Impact of Tannery on the worker's Reproductive Health in the area of Hazaribagh in Dhaka, Bangladesh. *International Journal of Research*, 2, pp. 159-181.

Dey, P., Raju, M.R. and Choudhury, M.R., 2016. Potential of phytoremediation approaches in the regeneration of Hazaribagh brownfield area. Proceedings of the 3rd International Conference on Civil Engineering for Sustainable Development. KUET, Khulna, Bangladesh.

Guilizzoni, P., 1991. The role of heavy metals and toxic metals and toxic materials in the physiological ecology of submerged macrophytes. *Aquatic Botany*,

41, pp. 87-109.

Haque, M.A., Chowdhury, R.A., Chowdhury, W.A., Baralaskar, A.H., Bhowmik, S. and Islam, S., 2019. Immobilization possibility of tannery wastewater contaminants in the tiles fixing mortars for eco-friendly land disposal. *Journal of environmental management*, 242, pp. 298–308.

Haque, S.E., Nahar, N., Chowdhury, S., Sakib, A.S., Saif, A., Hasan, S., Gomes, A.S. and Nezum, S.T., 2020. Impacts of the partial relocation of Hazaribagh tanneries on the environment and human health: Focus on children and vulnerable population. *International Journal of Students' Research in Technology and Management*, 8(4), pp. 1-7.

Hasan, M.M., Hosain, S., Asaduzzaman, A.M., Haque, M.A. and Roy, U.K., 2016. Prevalence of health diseases among Bangladeshi tannery workers and associated risk factors with workplace investigation. *Journal of Pollution Effects & Control*, 4, pp. 175.

Hasan, S.M.M., Akber, M.A. and Bahar, M.M., 2021. Chromium contamination from tanning industries and phytoremediation potential of native plants: A study of Savar tannery industrial estate in Dhaka, Bangladesh. *Bulletin of Environmental Contamination and Toxicology*, 106, pp. 1024–1032.

Hashem, M.A., Islam, A., Paul, A. and Nasrin, S., 2014. Generation of ammonia in deliming operation from tannery and its environmental effect: Bangladesh perspective. *International Journal of Energy and Environmental Engineering*, 2(4), pp. 266-270.

Hossain, A.M., Monir, T., Ul-Haque, A.R., Kazi, M.A.I., Islam, M.S. and Elahi, S.F., 2007. Heavy metal concentration in tannery solid wastes used as poultry feed and the ecotoxicological consequences. *Bangladesh Journal of Scientific and Industrial Research*, 42(4), pp. 397–416.

International Labor Organization. 2008. Decent work country program, Bangladesh. International Labour Office, Geneva.

Islam, G.M.R., Khan, F.E., Hoque, M.M. and Jolly, Y.N., 2014. Consumption of unsafe food in the adjacent area of Hazaribag tannery campus and Buriganga River embankments of Bangladesh: heavy metal contamination. *Environmental Monitoring and Assessment*, 186(11), pp. 7233-7244.

Islam, M.M., Karim, M.R., Zheng, X. and Li, X., 2018. Heavy metal and metalloid pollution of soil, water and foods in Bangladesh: A critical review. *International Journal of Environmental Research and Public Health,* 15(12), pp. 2825.

Juel, M.A., Chowdhury, Z.U., Mizan, A. and Alam, M.S., 2016. Toxicity and environmental impact assessment of heavy metals contaminated soil of Hazaribagh tannery area. Proceedings of 3rd International Conference on Advances in Civil Engineering. Chittagong: CUET.

Kamran, M.A., Amna, Mufti, R., Mubariz, N., Syed, J.H., Bano, A., Javed, M.T., Munis, M.F.H., Tan, Z. and Chaudhary, H.J., 2014. The potential of the flora from different regions of Pakistan in phytoremediation: a review. *Environmental Science and Pollution Research*, 21, pp. 801–812.

Kanagaraj, J., Velappan, K.C., Babu, N.K.C. and Sadulla, S., 2006. Solid wastes generation in the leather industry and its utilization for a cleaner environment: a review. *Journal of Scientific & Industrial Research*.65, pp. 541-548.

Khan, S., Cao, Q., Zheng, Y.M., Huang, Y.Z. and Zhu, Y.G., 2008. Health risks of heavy metals in contaminated soils and food crops irrigated with wastewater in Beijing, China. *Environ Pollution*, 152(3), pp. 686-92.

Laws of Brunei. (2001). Public health (food) regulations.

Mizan, A., Juel M.A.I., Alam M.S., Pichtel J. and Ahmed, T., 2020. Environmental and health risks of metal-contaminated soil in the former tannery area of Hazaribagh, Dhaka. *SN Applied Sciences*, 2, 1915. https://doi.org/10.1007/s42452-020-03680-4

Mondol, M.N., Asia, A., Chamon, A.S. and Faiz, S.M.A., 2017. Contamination of soil and plant by the Hazaribagh tannery industries. *Journal of the Asiatic Society of Bangladesh, Science,* 43(2), pp. 207–222.

Rahaman, A., Afroze, J.S., Bashar, K., Ali, M.F. and Hosen, M.R., 2016. A comparative study of heavy metal concentration in different layers of tannery vicinity soil and near agricultural soil. *American Journal of Analytical Chemistry*, 7, pp. 880-889.

Shahzad, K., Akhtar, S. and Mahmud, S., 2006. Prevalence and determinants of asthma in adult male leather tannery workers in Karachi, Pakistan: A cross-sectional study. *BMC Public Health*, 6, pp. 292.

Sultana, M.S., Jolly, Y.N., Yeasmin, S., Islam, A., Sattar, S. and Tareq, S.M., 2015. Transfer of heavy metals and radionuclides from soil to vegetables and plants in Bangladesh. *Soil Remediation and Plants*, pp. 331-366.

Thakur, L.S. and Parmar, M., 2013. Adsorption of heavy metal (Cu^{2+}, Ni^{2+}, and Zn^{2+}) from synthetic wastewater by tea waste adsorbent adsorption of heavy metal (Cu^{2+}, Ni^{2+}, and Zn^{2+}) from synthetic wastewater by Tea Waste Adsorbent. *International Journal of Chemical and Physical*

Sciences, 2, pp. 6–19.

Tinni, S.H., Islam, M., Fatima, K. and Ali, M.A., 2014. Impact of tanneries waste disposal on the environment in some selected areas of Dhaka City Corporation. *Journal of Environmental Science and Natural Resources*, 7(1), pp. 149-156.

Verma, R., Vijayalakshmy, K. and Chaudhiry, V., 2018. Detrimental impacts of heavy metals on animal reproduction: A review. *Journal of Entomology & Zoology Studies*, 6, pp. 27-30.

WHO (World Health Organization). 2004. The permissible level of metal in fish for human consumption.

WHO. 1984. The permissible level of metal in vegetables for human consumption.

WHO. 1996. Trace Elements in Human Nutrition and Health, Geneva.

Zahid, A., Balke, K.D., Hassan, M.Q. and Flegr, M., 2006. Evaluation of aquifer environment under Hazaribagh leather processing zone of Dhaka city. *Environmental Geology*, 50(4), pp. 495–504.

Journal of FST

VOLUME-01 | ISSUE-01 | JULY 2022
ISSN: 2959-4812

Bangladesh University of Professionals (BUP)
Mirpur Cantonment, Dhaka-1216
Telephone : 88-02-8000368, Fax : 88-02-8000443
E-mail: journal.fst@bup.edu.bd