International Conference on Emerging Technologies (ICET) 2022

DocOnTap: AI-based disease diagnostic system and recommendation system

1st Zain ul Abideen
Faculty of Computer Science & Engg.
AI Research Group
GIK Institute of Engg. Sciences & Tech.
Topi, Khyber Pakhtunkhwa, Pakistan.
zain-ul-abideen@giki.edu.pk

2nd Talha Ali Khan
Dept. of Tech & Software Engg.
Univ. of Europe of Applied Sciences
Berlin, Germany.
talhaali.khan@ue-germany.de

3rd Raja Hashim Ali
Faculty of Computer Science & Engg.
AI Research Group
GIK Institute of Engg. Sciences & Tech.
Topi, Khyber Pakhtunkhwa, Pakistan.
hashim.ali@giki.edu.pk

4th Nisar Ali
Faculty of Electronic Systems Engg.
University of Regina
Regina, Canada
nay095@uregina.ca

5th Muhammad Muneeb Baig
Faculty of Computer Science & Engg.
AI Research Group
GIK Institute of Engg. Sciences & Tech.
Topi, Khyber Pakhtunkhwa, Pakistan.
muneeb.baig@giki.edu.pk

6th Muhammad Sajid Ali
Faculty of Computer Science & Engg.
AI Research Group
GIK Institute of Engg. Sciences & Tech.
Topi, Khyber Pakhtunkhwa, Pakistan.
sajid.ali@giki.edu.pk

2022 17th International Conference on Emerging Technologies (ICET) | 978-1-6654-5992-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICET56601.2022.10004621

***Abstract*—In this age of technology, Artificial Intelligence plays a key part in humankind's growth, whether in education, daily life, or professional life. AI has improved how humans live and solve problems. Using illness mapping from symptoms, the suggested method recommends relevant doctors to users. The approach attempts to boost diagnosis efficiency, reducing misdiagnoses and saving doctors' time. Machine learning algorithms and doctors' diagnoses will reduce misdiagnosis. In this work, we built a disease-based prediction system with multiple machine learning algorithms including Decision Tree, Logistic Regression and Random Forest. We obtain the highest accuracy with Random Forest classifier.After diagnosis, our system will immediately schedule an appointment for them with the most conveniently located doctor in their area, and the system's evaluation will be delivered to the appointment doctor. The proposed system is accessible through website and both doctor and patient can use then for their purposes.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

In a report from the World Health Organization, it says that Pakistanis have trouble getting health care because they don't have enough money. More than 60 million people live below the poverty line and can't afford medical care. Even if they could pay for it, they can't afford to be wrongly diagnosed, which is another big problem in the medical field.The absence of accessible, high-quality healthcare is the root cause of the tragic events that have devastated so many lives and cut them tragically short. If over 197 million Pakistanis lack access to affordable, high-quality healthcare, they will be unable to contribute to the expansion of our economy. These statistics on health show that Pakistan's primary issue right now is not terrorism, drones, or the energy crisis, but rather the lack of sufficient medical treatment. Despite the fact that all of healthcare's problems are tied to corruption and a lack of cash, there is no reason why Pakistan's health sector shouldn't be its top priority right now. By doing this we are improving the current condition of healthcare sector in Pakistan. We aim at focusing on a very specific aspect of healthcare which is diagnosis of a disease. Doctors often misdiagnose a disease, which resultantly causes life threatening conditions. In a study, misdiagnosis rate was founded out to be 40%, which is quite high.

So, numerous treatments are available for various diseases, but there is no place where a person can get the detail of the diseases and their treatments [1], [2]. But in modern era, the E-Healthcare framework is an end client bolster, and this system enables clients to get indication on their disease and also drugs related to that disease [3], [4]. According to the World Health Organization e-Health signifies "the utilization of data and correspondence advancements ("ICT") for well being".

Our suggested E-Healthcare system intends to establish a better diagnostic system that is both transparent and convenient for the general population while requiring minimal work on their part. The major motivation is to reduce the rate of misdiagnosis and save time [5], [6]. We give a differentiated diagnostic system that eliminates all worries about the legitimacy and integrity of the results produced by implementing machine learning and AI approaches. Following disease prediction based on the symptoms supplied by the patients, our system will forward the evaluation to the doctor with whom the appointment has been set. This approach assists doctors in expanding their knowledge base while also saving them time. Our doctors do not have an excessive amount of time because they have to check numerous patients per day and

inquiring about each patient's initial history at each visit is time demanding for them. Our system will save the patient's history as well as assessments, making it easier for doctors to evaluate them at any time.

The proposed system automates appointment scheduling while saving the patient time [7], [8]. Patients can arrange appointments online and visit the doctor at the scheduled time, rather than waiting in huge lines. We tend to introduce a doctor plus technology solution. Using human intelligence plus technological innovations can provide better diagnosis and leading to a decrease in misdiagnosis rate [9], [10]. Providing a better healthcare standard to the common people of Pakistan. Upon the implementation of this system, it will be insured that patients are diagnosed correctly and treated well [11], [12]. Also, a full-fledged working website is developed and is available for both doctors and patients.

II. Literature Review

A lot of prior work has been done by computer scientists and numerous diagnosis apps and knowledge base apps have been developed for doctors.

Medical errors occur on a daily basis in the world [13]. The most prominent errors are error in diagnosis of a disease, providing a wrong drug, faulty record keeping, etc. 1.3 million people are injured due to the medication and misdiagnosis errors. A study carried out in Karachi revealed an error of 39.28 percent. The major reason behind such circumstances is the lack of accountability. There is a lack of a proper career structure for doctors which leads them to work constantly for hours resulting in errors and misdiagnosis. In accordance with a local newspaper the 8th leading cause of death in Pakistan is due to misdiagnosis. For physicians managing 1500 to 2000 patients, AI is an opportunity to help them.

Medical diagnosis apps are such apps that are used to diagnose illness a person might have based on the data gathered through a series of questions and answers. A recent study states that almost 42 percent doctors believed their patients can benefit from online apps. Such apps can aid in diagnosis for the doctors as well. [14].

Epocrates is the most downloaded medical application that provides a lot of medical benefits. It diagnosis diseases and also describes effect and side effects of a prescribed drug through a person's BMI.

PEPID's Symptom Checker, diagnosis diseases through patients' symptoms, lab reports and physical exam findings. It also provides prescription and dosage all in one screen. But unfortunately, it is a paid application not an open resource.

Symptomate is a website based on infermedica api that is responsible for disease prediction. It diagnoses diseases based on cross questioning from the patient and then give certain predictions of diseases.

Google Health is an open source health aiding resource developed by the Google community. Google Health is working on the use of artificial intelligence to assist in diagnosing cancer, predicting patient results, blindness prevention and much more. They work in collab with a community of doctors, nurses and patients to provide better and improved health experiences.

Advance Medical is another such application that saves money as well as lives through it's medical services. It is stated that Advance Medical found saving of 10,000 per person, 29 percent of patients had less intense therapy, 15 percent reduction in patient's medication. The work of Advance Medical is aimed at decreasing the cost of medicine. All in all, the upsurge through technology can bring about huge transformation in medicine and portal like Advance Medical will help medical segments. Hence saving lives through technology will save billions of dollars [15].

An update has been recently added to Google in which Google plays doctor by identifying your medical symptoms. One can easily ask Google for medical advices and also search for symptoms. Instead of searching for a whole condition a person can search for a certain symptom like 'my head hurts' and Google will provide conditions, treatments and home remedies about it. Although this feature will massively benefit the general public, Google still urges the fact that the person should seek a medical advice as a precaution and in case of serious situation in which the ailment itself is deadly [16].

According to a survey with a response rate of 58.6% (140 out of 240 participants), 90.7% of the participants were aware of different health applications and 50.1% of the people had installed health applications on their phones. Hence the regular use of mobile phones is pretty common in students and the study revealed that use of medical apps is useful to track physical activity and calorie intake as well as body mass index (BMI) [17].

They provides an API to prescribe medications to users with a specific ailment, which the framework diagnoses by studying the user's symptoms using machine learning techniques. They use data mining to determine the most likely illness based on symptoms. The patient can recognise diseases. Patients can identify their illness by describing their symptoms, and the app's interface displays the ailment [18]. The author aims to solve the problem by recommending treatments based on a user's symptoms. This research suggests a mechanism to reduce the laborious procedure of visiting hospitals (especially during a pandemic) and receiving a doctor's appointment. The author applies Natural language processing and machine learning to construct a chatbot app. The chatbot application will try to anticipate the disease and offer treatment through a series of questions. People will be able to do daily health check-ups, it will make them aware of their health status, and it will urge them to take correct actions to be healthy for free [19].

III. Methodology

For this research, we extracted data from the Ada Health Care website and the Infermedica API. The data is then preprocessed by getting rid of any unnecessary characters including emoticons, links, numbers, empty rows, and spaces. Then, we use word2vec on the possessed data and we implemented three different machine learning classifiers: Decision Tree, Logistic

Regression, and Random Forest. The complete methodology and workflow of our contribution are discussed in Figure 1.

Initially, there were many attempts to gather information from Pakistani medical facilities. Data gathering was aided by emails sent to hospital administration and by a series of meetings set up with hospital administration.

The hospitals all gave the same excuse—that patient information is confidential and therefore cannot be released to the public. This is why we used the Infermedica API to get information on diseases and their symptoms. The API can be used to find a correlation between symptoms and diseases. All queries and answers from the API are handled in the JSON format. While the API is not free, a student membership was used to keep costs down. Taking it into the real world, though, will necessitate a paid subscription.There are only 687 diseases that can be mapped using the Infermedica API. The diseases mapped by the API are all fairly common ones; it does not map every possible disease. We do preprocessing because, despite the fact that the data is already in a preprocessed state, we discovered during our work that more cleaning was necessary.

Information on a disease's signs and precautions is essential for the patient. As a result, we need additional data to analyze and train the model. So, using python and beautiful soup, we extracted information about diseases from the Ada Health Care website. Since the data require some preproceesing, we apply preprocessing to get them ready. This information was then incorporated into the online app so that patients, after receiving the prediction result, can learn more about their respective ailments.

After the data collections, a word2vec model was trained to produce disease embeddings. A word2vec model is a 2-layer neural network usually used for producing context of words but in the project, it was used for producing disease vectors [20]. The mapping of symptoms to diseases represented data that was fed into the model, which ultimately resulted in the production of a vector space. In vector space, disorders that shared symptoms were found to be clustered closely together [21]. It is easier to diagnose diseases with similar symptoms if the vector space of each disease is known. Natural language processing (NLP) techniques like text classification were also tried to identify diseases from symptoms, however they yielded unsatisfactory results for diseases with similar signs and symptoms.

A. Decision Trees

Decision Trees is supervised learning algorithm. The decision tree technique can help address regression and classification problems [22], [23]. Decision Trees are used to develop a training model that can predict the class or value of a target variable from prior data (training data). Start at the tree's root to forecast a record's class label. Compare root and record attributes. Using the comparison, we follow the branch to the next node. Decision trees classify examples by descending the tree from the root to a leaf/terminal node. Each node in the tree is a test case for some property, and each edge is a possible answer. Each new node-rooted subtree repeats this process.

B. Logistic Regression

Logistic Regression is a prominent supervised AI method. Logistic Regression is a probability-based statistical approach and Machine Learning algorithm for classification problems [24]. If the dependent variable (target) is categorical, use this technique. It is utilised when the classification problem is binary, such as true or false, yes or no. It can tell if an email is spam or not. Logistic regression uses the sigmoid function to calculate label likelihood. By rule, logistic regression values must be between 0 and 1. It produces a "S" curve on a graph due to its 1 limit. Here's how to calculate the sigmoid or logistic function. The threshold is used in Logistic Regression.

The working of the Logistic Regression model is shown schematically in Figure 2.

C. Random forest

Among the many supervised machine learning algorithms, random forest is widely used for classification and regression tasks. It constructs decision trees from the collected data and uses majority vote classification to assign labels to the samples. One of the most useful aspects of the Random Forest Algorithm is that it can be used for regression and classification, which involve data sets with both continuous and categorical variables. It achieves better results than competing algorithms for classification problems [25]. Some of the procedures in a random forest include:

- In a random forest, n records are chosen at random from a set of k records.
- With each sample, a new decision tree is developed.
- Each decision tree yields a result.
- When it comes to classification, the result that stands in the end is decided by majority vote.

Figure 3 illustrates the working of the Random Forest algorithm on the test dataset.

IV. Results and Discussions

This research evaluates and contrasts the efficacy of different classifiers on the dataset in order to improve disease prediction. The effectiveness of various classifiers was evaluated with the use of accuracy evaluation metrics. We then ran many classifiers on Infermedica API data and Infermedica API data with Ada Data that had been scraped.

A. Accuracy

Accuracy is the ratio of a correct prediction made by the classifier to the total prediction made by the classifier

$$Accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

The random forest outperforms the other classifiers on both sets of data (Infermedica API and Infermedica API + Ada Web Data), with an accuracy of 80.23% and 87.23%, respectively. Infermedica API data and Infermedica API data combined with data scraped from the Ada Web are used to illustrate the accuracy scores of several classifiers (Figure 4).

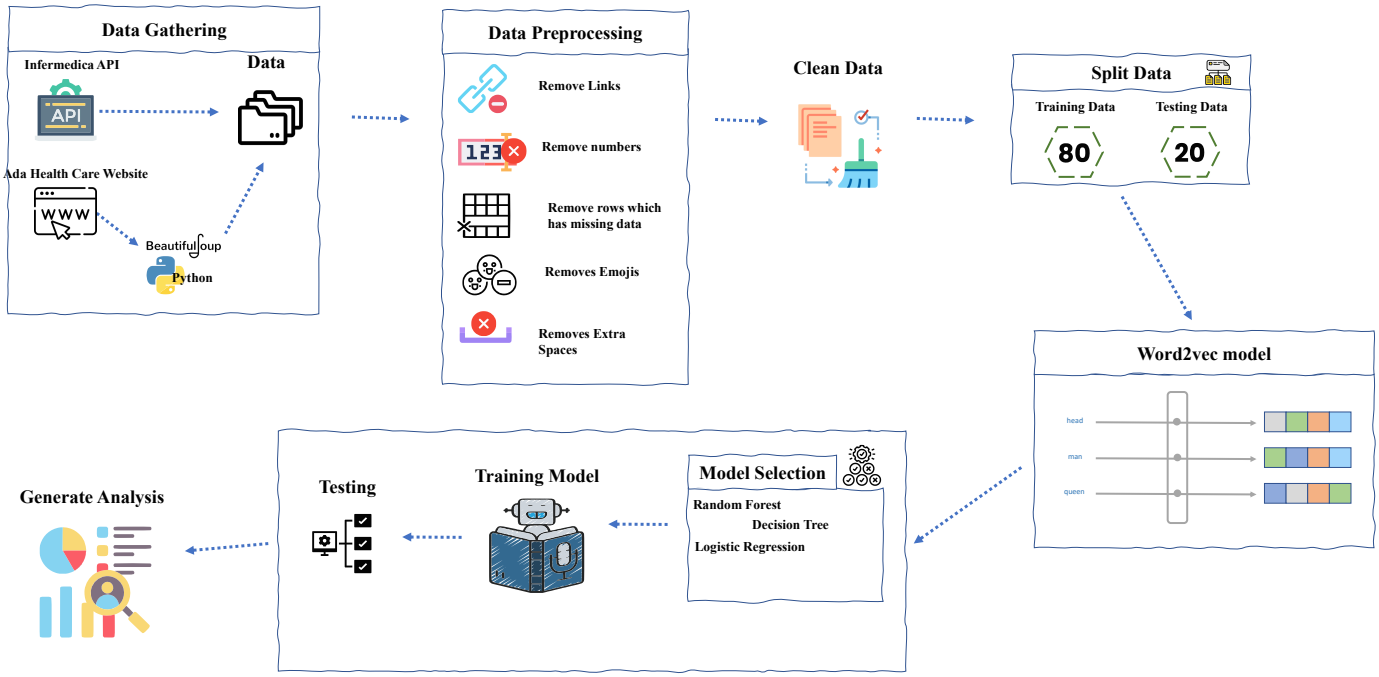

Fig. 1. The overall working of the proposed solution

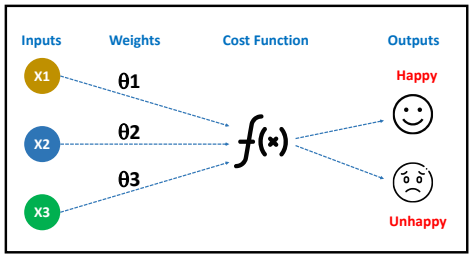

Fig. 2. An illustration of how the Logistic Regression algorithm works as a classification technique

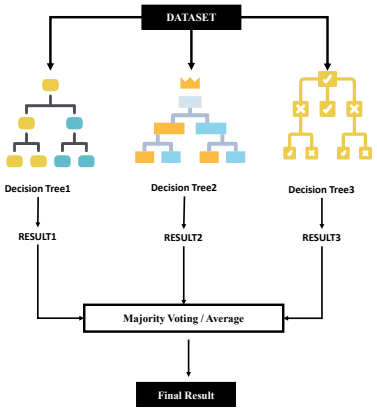

Fig. 3. An illustration of how Random Forest works as a classification technique

The accuracy of the models for a prediction phase was with decision tree, logistic regression and random forest 67.25%, 72.67% and 80.23% respectively. The accuracy is quite low given the fact that the problem was related to a health field where accuracy is quite essential. But since the symptoms were randomized a little to get three different disease probabilities.

The accuracy of the models increases to 70%, 80%, and 87.23% when we integrate the data from the Infermedica API with the data from the Ada Health Care website and then re-train the models to verify their performance.

(((By combining human and machine accuracy, the system may then schedule an appointment with the doctor, increasing the likelihood of an accurate diagnosis and reducing the misdiagnosis rate.)))

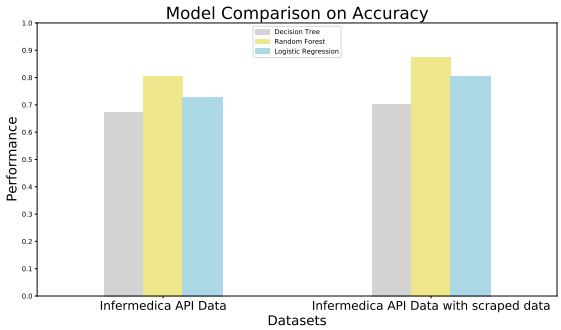

Fig. 4. The overall working of the proposed solution

V. Conclusion and Future Work

In today's busy society, individuals are too busy to go to the doctor when they feel sick, which can be fatal and lead to critical sickness. World misdiagnosis rate is 40%, which is high. DocOnTap, a web-based software, lets patients examine their medical issues. The patient can assess anyplace with a gadget and internet connection. After the assessment, the patient can book an appointment with a doctor, who can see system disease predictions. Combining human and machine accuracy can reduce misdiagnosis and improve human life.

In traditional medical care, the doctor asks about the patient's medical history to diagnose them. DocOnTap saves time for both the patient and doctor by providing access to medical history when the patient makes an appointment. Since the doctor has the patient's complete medical history, it saves time and improves accuracy. The application's differential diagnosis procedure makes user involvement very interactive, and the patient finds it no different than a doctor.

Those who understand medical jargon and English can use the system effectively. When a user selects an explain button, medical terms will be explained. Future updates will add numerous language support so people from all backgrounds can use the software. Developing an Android and IOS app will make the system more accessible. The app is user-friendly. In today's society, customers prefer mobile apps to online apps on phones.

At present, we are only able to train three models, and those models are only being used to look for common diseases. Adding more features and employing feature engineering will improve the performance of the system, allowing for more precise disease diagnosis. By gathering data from several APIs, preprocessing it, and combining it into a single platform, the system may be made to handle a much larger number of diseases. Thereafter, the models will need to be retrained so that the system can treat a wide variety of ailments. Furthermore, we want to implement deep learning techniques in order to boost the system's overall efficiency. On top of that, we'll develop a mobile app for it.

Acknowledgment

We are thankful to Dean Faculty of Computer Science and Engineering at GIK Institute, Dr. Ahmar Rashid, for providing us support in carrying out this research and GIK Institute for providing us a platform.

References

[1] R. H. Ali, S. A. Muhammad, and L. Arvestad, "Genfamclust: an accurate, synteny-aware and reliable homology inference algorithm," *BMC evolutionary biology*, vol. 16, no. 1, pp. 1–19, 2016.

[2] R. H. Ali, S. A. Muhammad, M. A. Khan, and L. Arvestad, "Quantitative synteny scoring improves homology inference and partitioning of gene families," in *BMC bioinformatics*, vol. 14, no. 15. BioMed Central, 2013, pp. 1–9.

[3] D. K. Mohammad, R. H. Ali, J. J. Turunen, B. F. Nore, and C. E. Smith, "B cell receptor activation predominantly regulates akt-mtorc1/2 substrates functionally related to rna processing," *PloS one*, vol. 11, no. 8, p. e0160255, 2016.

[4] M. Mushtaq, R. H. Ali, V. Kashuba, G. Klein, and E. Kashuba, "S18 family of mitochondrial ribosomal proteins: evolutionary history and gly132 polymorphism in colon carcinoma," *Oncotarget*, vol. 7, no. 34, p. 55649, 2016.

[5] R. H. Ali, M. Bark, J. Miró, S. A. Muhammad, J. Sjöstrand, S. M. Zubair, R. M. Abbas, and L. Arvestad, "Vmcmc: a graphical and statistical analysis tool for markov chain monte carlo traces," *BMC bioinformatics*, vol. 18, no. 1, pp. 1–8, 2017.

[6] R. H. Ali, M. Bogusz, and S. Whelan, "Identifying clusters of high confidence homologies in multiple sequence alignments," *Molecular biology and evolution*, vol. 36, no. 10, pp. 2340–2351, 2019.

[7] A. A. Khan, R. H. Ali, and B. Mirza, "Evolutionary history of alzheimer disease-causing protein family presenilins with pathological implications," *Journal of Molecular Evolution*, vol. 88, no. 8, pp. 674–688, 2020.

[8] Y. Yang, S. Tu, R. H. Ali, H. Alasmary, M. Waqas, and M. N. Amjad, "Intrusion detection based on bidirectional long short-term memory with attention mechanism," *CMC – Computer Material and Continua*, vol. 74, no. 1, pp. 1597–1632, 2022.

[9] I. Ul Hassan, R. H. Ali, Z. Ul Abideen, T. A. Khan, and R. Kouatly, "Significance of machine learning for detection of malicious websites on an unbalanced dataset," *Digital*, vol. 2, no. 4, pp. 501–519, 2022.

[10] Z. Halim, S. Hussain, and R. H. Ali, "Identifying content unaware features influencing popularity of videos on youtube: A study based on seven regions," *Expert Systems with Applications*, vol. 206, p. 117836, 2022.

[11] R. H. Ali, "From genomes to post-processing of bayesian inference of phylogeny," Ph.D. dissertation, KTH Royal Institute of Technology, 2016.

[12] H. R. Ali, *Determining the relationship between gene class network graph and gene age*. Chalmers University of Technology, 2009.

[13] L. E. Søvold, J. A. Naslund, A. A. Kousoulis, S. Saxena, M. W. Qoronfleh, C. Grobler, and L. Münter, "Prioritizing the mental health and well-being of healthcare workers: an urgent global public health priority," *Frontiers in public health*, vol. 9, p. 679397, 2021.

[14] J. C. Moses, S. Adibi, S. M. Shariful Islam, N. Wickramasinghe, and L. Nguyen, "Application of smartphone technologies in disease monitoring: a systematic review," in *Healthcare*, vol. 9, no. 7. MDPI, 2021, p. 889.

[15] P. Agyemang-Gyau, "Artificial intelligence in healthcare and the implications for providers," *On-Line Journal of Nursing Informatics*, vol. 25, no. 2, 2021.

[16] T. Alanzi, "A review of mobile applications available in the app and google play stores used during the covid-19 outbreak," *Journal of multidisciplinary healthcare*, vol. 14, p. 45, 2021.

[17] G. Singh and S. Alva, "A survey on usage of mobile health apps among medical undergraduates," *HSOA Journal of Community Medicine and Public Health Care*, vol. 6, no. 1, pp. 1–6, 2019.

[18] V. Mudaliar, P. Savaridaasan, and S. Garg, "Disease prediction and drug recommendation android application using data mining (virtual doctor)," *International Journal of Recent Technology and Engineering*, vol. 8, 2019.

[19] J. Agarwal, M. Kumar, and A. K. Srivastava, "Symptoms based disease diagnosis and treatment recommendation," in *2021 International Conference on Computational Performance Evaluation (ComPE)*. IEEE, 2021, pp. 162–167.

[20] A. Khatua, A. Khatua, and E. Cambria, "A tale of two epidemics: Contextual word2vec for classifying twitter streams during outbreaks," *Information Processing & Management*, vol. 56, no. 1, pp. 247–257, 2019.
[21] J. Sarzynska-Wawer, A. Wawer, A. Pawlak, J. Szymanowska, I. Stefaniak, M. Jarkiewicz, and L. Okruszek, "Detecting formal thought disorder by deep contextualized word representations," *Psychiatry Research*, vol. 304, p. 114135, 2021.
[22] M. M. Ahsan, S. A. Luna, and Z. Siddique, "Machine-learning-based disease diagnosis: A comprehensive review," in *Healthcare*, vol. 10, no. 3. MDPI, 2022, p. 541.
[23] N. Ali, S. Ansari, Z. Halim, R. H. Ali, M. F. Khan, and M. Khan, "Breast cancer classification and proof of key artificial neural network terminologies," in *2019 13th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics (MACS)*. IEEE, 2019, pp. 1–6.
[24] I. Ibrahim and A. Abdulazeez, "The role of machine learning algorithms for diagnosing diseases," *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 10–19, 2021.
[25] S. Asadi, S. Roshan, and M. W. Kattan, "Random forest swarm optimization-based for heart diseases diagnosis," *Journal of Biomedical Informatics*, vol. 115, p. 103690, 2021.