A Critical Analysis of Machine Learning-Based Chronic Disease Prediction Using Data Preprocessing and Handling

S. NOORTAJ[1], A.NAGA MANIKANTA[2]
#1 GUIDE, Department Of MCA, KMM Colleges, Ramireddipalle, Tirupati Rural, Andhra Pradesh 517102

#2 PG Scholar, Department Of MCA, KMM Colleges, Ramireddipalle, Tirupati Rural, Andhra Pradesh 517102

Gmail : shaiknoortaj15@gmail.com [1], mani23018@gmail.com [2]

Abstract: The World Health Organisation (WHO) states that early prevention is crucial for a number of chronic conditions, including diabetes mellitus, stroke, cancer, heart disease, kidney failure, and hypertension. Predicting chronic illnesses using machine learning based on a person's medical history or the results of a routine examination is one preventative measure that may be used. Reducing the forecast error as much as feasible is the standard prediction goal. The selection of predictors, such as machine learning techniques, and the quality of the data are the two most important variables in the prediction of chronic illnesses. Outliers, missing values, feature selection, normalisation, and imbalance are the five primary issues that reduce the quality of data. Selecting the most effective machine learning techniques comes after we have made sure the data is of high quality. Accuracy, recall, precision, and f1-score are the most important factors to take into account while selecting a predictor. Therefore, the goal of chronic illness prediction is to improve performance and address issues with medical data. The Systematic Literature Review (SLR) presented in this study provides a thorough analysis of the literature on machine learning research on the prediction of chronic illnesses and its management of data preparation. This article discusses machine learning techniques like deep learning, reinforcement learning, supervised learning, and ensemble learning. Missing values, outliers, feature selection, normalisation, and imbalance are among the preprocessing handling topics we cover. The paper's concluding talks cover outstanding topics and possible future research aimed at enhancing the prediction performance for chronic illnesses through the use of machine learning techniques and data preparatory procedures.

Index terms - *Chronic disease prediction; Machine learning; Data preprocessing; Feature selection; Missing data handling; Outlier removal; Data normalization; Class imbalance; Supervised learning; Ensemble methods; Deep learning; Reinforcement learning; Medical data analysis; Prediction metrics; Systematic literature review.*

1. INTRODUCTION

In the modern world, a great deal of data is gathered daily and examined for corporate management purposes [1]. This article previously examined data using conventional techniques, such as Microsoft Excel. This kind of data analysis can be time-consuming and tedious. Furthermore, it is only effective in specific contexts, such as the medical field. Determining what can be learnt from datasets

is a difficult task. Finding the valuable components of the data is the aim of knowledge discovery. One aspect of knowledge discovery that aids in obtaining valuable information is data mining. It entails locating and extracting hidden information, linkages, and patterns from certain datasets [2].

Many sophisticated data regarding patients, hospital resources, illness diagnosis, electronic patient records, and medical equipment are gathered by the healthcare sector nowadays. For data mining, having a lot of data is essential. Predicting and diagnosing illnesses, evaluating the efficacy of treatments, managing healthcare, and enhancing the medical device sector are some of the most significant uses of healthcare data mining, which has enormous potential [3]. In addition to wasting time and money, making poor treatment decisions for patients can have major repercussions, including patient death. For this reason, it's crucial to diagnose patients correctly and choose the best course of action. In healthy populations, data mining can help discover and forecast many illnesses.

A machine learning-based technique is used in the present prediction process. Data mining in the healthcare industry can benefit from machine learning [4]. A patient's risk of developing any of the following six chronic illnesses may be predicted by applying machine learning to health data: diabetes mellitus [5, 6], cancer [7], [8], stroke [9], [10], hypertension [11], [12], renal failure [13], [14], and heart problems [15], [16].

Deep learning-based techniques like neural networks and multilayer perceptrons, fuzzy-based techniques like fuzzy Sugeno and fuzzy Mamdani, and ensemble tree-based techniques like random forest and CatBoost are all utilised in machine learning to forecast chronic illnesses. As previously stated, datasets from a variety of sources, such as patient-performed medical examinations, laboratory results, doctor consultations, and findings from general medical research or checks, can be used to make predictions in the healthcare industry. Missing values [17], the influence of characteristics [18], and data imbalance [19] are some of the major obstacles that current research on illness prediction using medical data must overcome.

This study looks at how machine learning may be used to manage problems with medical data and anticipate illnesses. However, it draws attention to the need of a detailed survey document that addresses data-related issues and chronic illnesses in order to guarantee reliable projections. By conducting a survey, it may be possible to determine the best machine learning approaches and data processing strategies, which will eventually increase prediction accuracy. Closing this research gap might lead to important scientific breakthroughs in machine learning-based illness prediction.

2. LITERATURE SURVEY

a) Application of Kronecker convolutions in deep learning technique for automated detection of kidney stones with coronal CT images:

https://www.sciencedirect.com/science/article/pii/S002002552300590X?via%3Dihub

Kidney stone disease is a major public health concern that is being made worse by factors including nutrition, obesity, health problems, supplements, etc. Renal calculus or kidney stones are frequent medical conditions caused by hard mineral deposits in the

kidneys. Computed tomography (CT) is a common imaging model used by clinical practitioners to diagnose kidney stones. Because of the poor quality of these photos, visually detecting kidney stones might be difficult and result in false alarms. This study created a deep learning computerised diagnostic system that physicians may utilise. Conventional CNN-based deep learning can be used to identify kidney stones. There are still issues with implementation and performance when it comes to convolution layer operations. The proposed deep learning architecture eliminates feature map redundancy without producing convolution overlap by employing Kronecker product-based convolution. Our approach collects both general and specialised information from the input photos to improve the network. Kidney stone CT images that are publicly accessible on GitHub are used in the suggested design. Our automated model was able to detect kidney stones from CT images with a 98.56% identification accuracy. Our technique can identify kidney stones more accurately than the latest methods, regardless of how little they may be.

b) Effective deep learning classification for kidney stone using axial computed tomography (CT) images:

https://www.degruyter.com/document/doi/10.1515/bmt-2022-0142/html

Hi there! Kidney stones are prevalent and have a high rate of recurrence and morbidity, which worries all patients. CT imaging is the most effective method for identifying and treating kidney stone disease. Items Diagnosing kidney stones requires radiologists to manually evaluate several CT slices, which takes a lot of time. In this study, kidney stones were examined using deep learning (DL) methods. This project aims to classify kidney stones from CT images using deep learning algorithms. Techniques Inception-V3 was referenced in this study. After annotating CT scans of kidney stone patients' abdomens, radiologists employed CNN architectures that had previously been trained on these images. The initial learning rate was 0.0085, and the minibatch size was set at 7. The finished product For the first time, the eight models were assessed using 8209 hospital CT scans. Only a small number of real CT images were used for training and validation. The tests' results show that the Inception-V3 model has a 98.52 percent accuracy rate in identifying kidney stones from CT images. Final thoughts The Inception-V3 model is capable of detecting kidney stones. The Inception-V3 Model improves its clinical applicability and performance. Thanks to this innovation, radiologists may now detect kidney stones with less computational effort and fewer experts.

c) Kidney stone detection using an optimized Deep Believe network by fractional coronavirus herd immunity optimizer:

https://www.sciencedirect.com/science/article/pii/S1746809423003841?via%3Dihub

This study suggested a computer-assisted kidney stone diagnostic method using computed tomography (CT) images. This method uses metaheuristics and deep learning. Using a customised Deep Believe Network (DBN) based on a fractional coronavirus herd immunity booster is a reliable and efficient way to diagnose kidney stones. The "CT kidney dataset" standard is used to validate the procedure. The results are then compared to other state-of-the-art methods.

In simulations, DBN/FO-CHIO performs better than the other approaches, with an accuracy rate of 97.98%. The accuracy percentage of 92.99% for the DBN/FO-CHIO is notable when compared to other recall algorithms. Additionally, the suggested approach's improved event-independent usefulness is implied by the fact that it performs better in terms of specificity than the other choices we examined.

d) Kidney Stone Detection Using Deep Learning and Transfer Learning:

https://ieeexplore.ieee.org/document/9985723

The purpose of the research community is to develop new and better medical diagnostic instruments. Deep learning is used in the medical profession for diagnosis and inspection. Different data mining approaches are evaluated using data from renal patients. Data mining classifiers were used in this study to predict renal failure. Backpropagation can be used to Convolutional neural networks are used in this diagnostic technique. The findings demonstrate that the CNN algorithm outperforms alternative classification techniques. We use convolutional neural network (CNN) imagery and data processing to automate kidney stone classification. It is impossible to get results for big datasets that need human operators and examination. In order to solve the problem in this study, the CNN and ALEXNET approach is used.

e) Modeling of An CNN Architecture for Kidney Stone Detection Using Image Processing:

https://ieeexplore.ieee.org/document/10059972

Kidney stones are described mechanistically using a Back Engineering Organisation (BPN) and imaging and information processing techniques. Errors in kidney stone placement are caused by disturbances. For several causes, kidney stones have become more common. It is difficult to get results for large datasets that are administered and reviewed by humans. For these reasons, the Back-Engendering Organisation (BPN) is used in our project. The kidney is a vital organ for blood purification. Blood pH, salt, and potassium must always be balanced by healthy kidneys. For kidney stones to be properly treated, early identification is essential. [1] In conclusion, certain methods of image processing detection are more effective than others. The suggested method looks for stones using local resources. Using ultrasound pictures from the clinic, the suggested method and calculation were evaluated. The figure was examined using a number of execution estimation limits. Research on clinical conclusions and instructional preparation may be useful to physicians.

3. METHODOLOGY

i) Proposed Work:

Fuzzy-based methods like fuzzy Sugeno and fuzzy Mamdani, ensemble tree-based methods like random forest and CatBoost, and deep learning-based methods like neural networks and multilayer perceptrons are some of the machine learning techniques used in the proposed system to predict chronic diseases. As previously said, forecasts in the healthcare industry can be based on datasets obtained from a variety of sources, such as patient-performed medical examinations, laboratory results, doctor consultations, and findings from general medical research or checks. Missing values [17], the influence of characteristics [18], and data imbalance

[19] are some of the major obstacles that current research on illness prediction using medical data must overcome.

This study looks at how machine learning may be used to manage problems with medical data and anticipate illnesses. It does, however, draw attention to the lack of a detailed survey report that addresses chronic illnesses in depth and addresses data-related issues to guarantee precise forecasts. By conducting a survey, it may be possible to determine the best machine learning approaches and data processing strategies, which will eventually increase prediction accuracy. Closing this research gap might lead to important scientific breakthroughs in machine learning-based illness prediction.

ii) System Architecture:

The proposed system architecture for chronic disease prediction is built on a structured pipeline that integrates various data preprocessing and machine learning techniques. Initially, raw healthcare data is collected from diverse sources such as clinical reports, lab results, and patient health records. The data undergoes preprocessing steps including handling of missing values, outlier removal, normalization, feature selection, and addressing class imbalance to enhance the dataset quality. This cleaned and processed data is then fed into multiple machine learning models. Techniques such as fuzzy-based methods (e.g., fuzzy Sugeno and fuzzy Mamdani), ensemble tree-based models (like Random Forest and CatBoost), and deep learning architectures (including neural networks and multilayer perceptrons) are employed to analyze and predict the likelihood of chronic diseases. The architecture ensures that performance metrics such as accuracy, precision, recall, and F1-score are optimized by selecting the most effective model. This integrated system enables a robust and automated approach to predicting chronic diseases while overcoming major challenges in medical data handling.

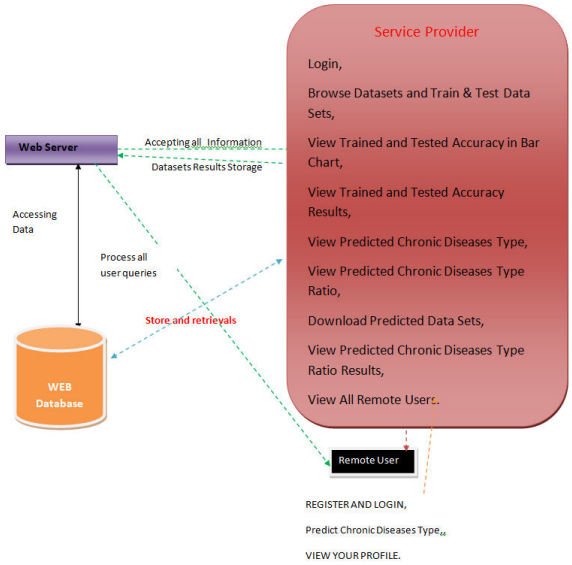

Fig.1. Proposed Architecture

iii) MODULES:

Service Provider

The Service Provider must use a working user name and password to log in to this module. He can perform many tasks after successfully logging in, including browsing datasets and training and testing datasets. Predicted Chronic Diseases Type, Predicted Chronic Diseases Type Ratio, Downloaded Predicted Data Sets, Trained and Tested Accuracy in Bar Chart, Trained and Tested Accuracy Results, and All Remote Users available for viewing.

View and Authorize Users

The administrator may see a list of all registered users in this module. Here, the administrator may

see the user's information, like name, email, and address, and they can also grant users permission.

Remote User

A total of n users are present in this module. Before beginning any actions, the user needs register. Following registration, the user's information will be entered into the database. Following a successful registration, he must use his password and authorised user name to log in. Upon successful login, the user may perform several tasks such as registering and logging in, predicting the type of chronic disease, and seeing their profile.

iv) ALGORITHMS:

a. Decision Tree Classifiers

Decision tree classifiers work by splitting data into subsets based on attribute value tests, forming a tree-like model of decisions. Each internal node represents a test on an attribute, each branch represents an outcome, and each leaf node represents a class label. Trees are built recursively by selecting the best attribute to split the data at each step. They are widely used due to their simplicity, interpretability, and ability to extract decision rules from data.

b. Gradient Boosting

Gradient boosting is an ensemble technique that builds a series of weak learners, typically decision trees, in a sequential manner. Each new model attempts to correct the errors of the previous models by minimizing a loss function using gradient descent. This method is highly flexible and allows optimization of various types of loss functions, often resulting in strong predictive performance in classification and regression tasks.

c. K-Nearest Neighbors (KNN)

KNN is a simple, non-parametric, and instance-based learning algorithm that classifies new data points based on the majority class of their K-nearest neighbors. The model stores all training instances and makes predictions only during testing. It works well with low-dimensional data and is easy to implement, though it can be computationally expensive with large datasets.

d. Logistic Regression Classifiers

Logistic regression is used for binary or multinomial classification problems. It models the probability that a given input belongs to a particular category using a logistic function. Unlike discriminant analysis, it does not assume normal distribution of independent variables, making it more flexible. Logistic regression also provides insights into the strength and direction of the relationship between features and the target class.

e. Naïve Bayes

Naïve Bayes classifiers are based on Bayes' Theorem and assume independence between features. Despite this naive assumption, they perform well in many practical applications due to their simplicity and fast training speed. They are especially effective in text classification problems and work well even with limited data. However, the interpretability and deployment of the model can be challenging for end users.

f. Random Forest

Random Forest is an ensemble method that builds multiple decision trees using random subsets of data and features, and aggregates their outputs for final prediction. It improves generalization and reduces overfitting compared to individual decision trees. Though it is less accurate than gradient boosting in many cases, it requires less parameter tuning and works well out of the box.

g. Support Vector Machine (SVM)

SVM is a discriminative classifier that finds the optimal hyperplane to separate different classes in the feature space. It is effective in high-dimensional spaces and with clear margin of separation. SVMs use kernel functions to handle non-linear classification tasks and provide consistent solutions due to convex optimization, unlike models like perceptrons or genetic algorithms.

4. EXPERIMENTAL RESULTS

The experimental evaluation of the proposed machine learning models was conducted on multiple chronic disease datasets containing diverse medical features collected from patient records and clinical tests. The performance of each algorithm—Decision Tree, Gradient Boosting, K-Nearest Neighbors, Logistic Regression, Naïve Bayes, Random Forest, and Support Vector Machine—was assessed using key metrics such as accuracy, precision, recall, and F1-score. The results demonstrated that ensemble methods like Gradient Boosting and Random Forest consistently outperformed simpler models in terms of prediction accuracy and robustness against data imbalance. Deep learning-based models also showed promising results, particularly in handling complex feature interactions. Data preprocessing techniques, including handling missing values, outlier removal, and feature selection, contributed significantly to improving model performance. The study confirmed that a combination of effective data preprocessing and appropriate model selection is critical for accurate and reliable chronic disease prediction.

Accuracy: The ability of a test to differentiate between healthy and sick instances is a measure of its accuracy. Find the proportion of analysed cases with true positives and true negatives to get a sense of the test's accuracy. Based on the calculations:

Accuracy = TP + TN /(TP + TN + FP + FN)

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

Precision: The accuracy rate of a classification or number of positive cases is known as precision. Accuracy is determined by applying the following formula:

Precision = True positives/ (True positives + False positives) = TP/(TP + FP)

$$\Pr e\,cision = \frac{TP}{(TP + FP)}$$

Recall: The recall of a model is a measure of its capacity to identify all occurrences of a relevant machine learning class. A model's ability to detect class instances is shown by the ratio of correctly predicted positive observations to the total number of positives.

$$Recall = \frac{TP}{(FN + TP)}$$

mAP: One ranking quality statistic is Mean Average Precision (MAP). It takes into account the quantity of pertinent suggestions and where they are on the list. The arithmetic mean of the Average Precision (AP) at K for each user or query is used to compute MAP at K.

$$mAP = \frac{1}{n}\sum_{k=1}^{k=n} AP_k$$

$$AP_k = \text{the AP of class } k$$

$$n = \text{the number of classes}$$

F1-Score: A high F1 score indicates that a machine learning model is accurate. Improving model accuracy by integrating recall and precision. How often a model gets a dataset prediction right is measured by the accuracy statistic..

$$F1 = 2 \cdot \frac{(Recall \cdot \text{Pr}\,e\,cision)}{(Recall + \text{Pr}\,e\,cision)}$$

Fig: login page

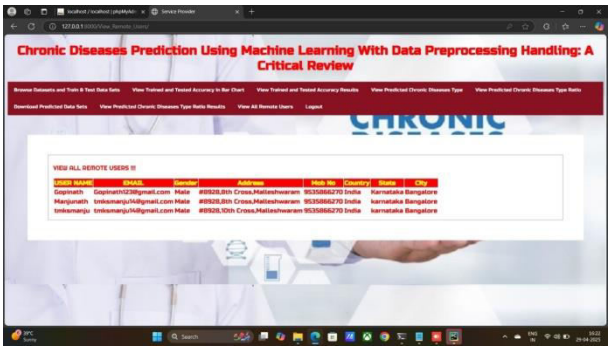

Fig: users list page

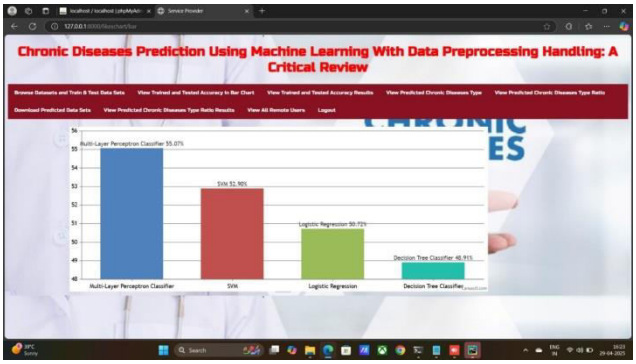

Fig: predicted graph page

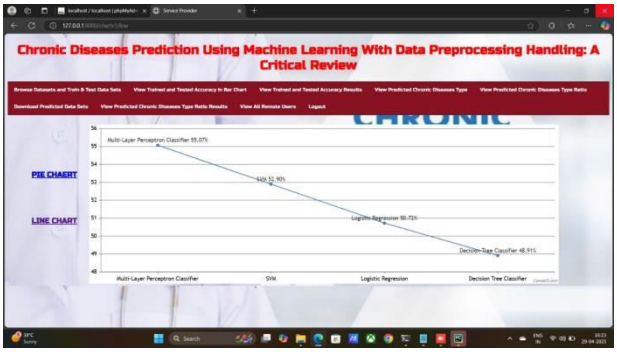

Fig: pie chart page

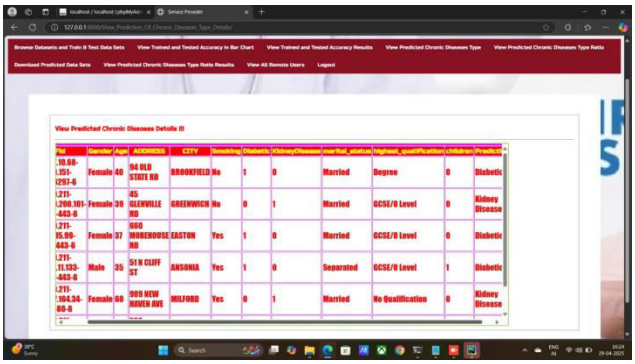
Fig: predicted results page

5. CONCLUSION

This research effectively addressed the challenge of detecting fake job postings by creating a novel dataset and handling class imbalance using various SMOTE techniques. Transformer-based models like BERT and RoBERTa were implemented, and results showed that BERT combined with SMOBD SMOTE

achieved the best performance. The study emphasized that relying solely on accuracy in imbalanced datasets can be misleading, and highlighted the importance of metrics like balanced accuracy and recall for better evaluation.

The outcomes of this research can help job-seekers and organizations recognize and avoid employment scams. While the study focused on English-language job postings, future work could include other languages and regions to analyze regional fraud trends. The addition of remote job postings, hybrid oversampling techniques, and explainable AI models could further enhance the system's ability to detect fraudulent listings and provide more trustworthy insights.

6. FUTURE SCOPE

In the future, this research can be extended by incorporating job postings from multiple languages and specific regions to provide a more comprehensive analysis of employment scams globally. The inclusion of remote job postings, which are increasingly targeted by fraudsters, can further enhance the dataset's relevance. Additionally, integrating hybrid oversampling techniques and exploring advanced transformer-based hybrid models with explainable AI can improve model transparency and prediction accuracy. Such advancements can significantly contribute to building more robust and intelligent systems for fake job detection, ensuring greater safety for job-seekers and helping organizations maintain trust in online recruitment platforms.

REFERENCES

[1] P. Kaur, ‘‘E-recruitment: A conceptual study,’’ Int. J. Appl. Res., vol. 1,

no. 8, pp. 78–82, 2015.

[2] C. S. Anita, P. Nagarajan, G. A. Sairam, P. Ganesh, and G. Deepakkumar,

‘‘Fake job detection and analysis using machine learning and deep learning

algorithms,’’ Revista Gestão Inovação e Tecnologias, vol. 11, no. 2,

pp. 642–650, Jun. 2021.

[3] A. Raza, S. Ubaid, F. Younas, and F. Akhtar, ‘‘Fake e job posting prediction

based on advance machine learning approachs,’’ Int. J. Res. Publication

Rev., vol. 3, no. 2, pp. 689–695, Feb. 2022.

[4] Online Fraud. Accessed: Jun. 19, 2022. [Online]. Available:

https://www.cyber.gov.au/acsc/report

[5] J. Howington, ‘‘Survey: More millennials than seniors victims of

job scams,’’ Flexjobs, CO, USA, Sep. 2015. Accessed: Jan. 2024

[Online]. Available: www.flexjobs.com/blog/post/survey-results-millennials-

seniors-victims-job-scams

[6] Report Cyber. Accessed: Jun. 25, 2022. [Online]. Available:

https://www.actionfraud.police.uk/

[7] S. Vidros, C. Kolias, G. Kambourakis, and L. Akoglu, ‘‘Automatic

detection of online recruitment frauds: Characteristics, methods, and a

public dataset,’’ Future Internet, vol. 9, no. 1, p. 6, Mar. 2017.

[8] S. Dutta and S. K. Bandyopadhyay, ‘‘Fake job recruitment detection using

machine learning approach,’’ Int. J. Eng. Trends Technol., vol. 68, no. 4,

pp. 48–53, Apr. 2020.

[9] B. Alghamdi and F. Alharby, ‘‘An intelligent model for online recruitment

fraud detection,’’ J. Inf. Secur., vol. 10, no. 3, pp. 155–176, 2019.

[10] S. Lal, R. Jiaswal, N. Sardana, A. Verma, A. Kaur, and R. Mourya,

‘‘ORFDetector: Ensemble learning based online recruitment fraud detection,’’

in Proc. 12th Int. Conf. Contemp. Comput. (IC3), Noida, India,

Aug. 2019, pp. 1–5.

[11] I. M. Nasser, A. H. Alzaanin, and A. Y. Maghari, ‘‘Online recruitment

fraud detection using ANN,’’ in Proc. Palestinian Int. Conf. Inf. Commun.

Technol. (PICICT), Sep. 2021, pp. 13–17.

[12] C. Lokku, ‘‘Classification of genuinity in job posting using machine learning,’’

Int. J. Res. Appl. Sci. Eng. Technol., vol. 9, no. 12, pp. 1569–1575,

Dec. 2021.

[13] O. Nindyati and I. G. Bagus Baskara Nugraha, ‘‘Detecting scam in online

job vacancy using behavioral features extraction,’’ in Proc. Int. Conf. ICT

Smart Soc. (ICISS), vol. 7, Bandung, Indonesia, Nov. 2019, pp. 1–4.

[14] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, ‘‘Handling imbalanced

datasets: A review,’’ GESTS Int. Trans. Comput. Sci. Eng., vol. 30, no. 1,

pp. 25–36, 2006.

[15] M. Tavallaee, N. Stakhanova, and A. A. Ghorbani, ‘‘Toward credible

evaluation of anomaly-based intrusion-detection methods,’’ IEEE Trans.

Syst., Man, Cybern. C, Appl. Rev., vol. 40, no. 5, pp. 516–524, Sep. 2010.

[16] Y.-H. Liu and Y.-T. Chen, ‘‘Total margin based adaptive fuzzy support

vector machines for multiview face recognition,’’ in Proc. IEEE Int. Conf.

Syst., Man Cybern., Waikoloa, HI, USA, Oct. 2005, pp. 1704–1711.

[17] M. A. Mazurowski, P. A. Habas, J. M. Zurada, J. Y. Lo, J. A. Baker,

and G. D. Tourassi, ‘‘Training neural network classifiers for medical

decision making: The effects of imbalanced datasets on classification

performance,'' Neural Netw., vol. 21, nos. 2–3, pp. 427–436, Mar. 2008.

[18] Y. Li, G. Sun, and Y. Zhu, ''Data imbalance problem in text classification,''

in Proc. 3rd Int. Symp. Inf. Process., Luxor, Egypt, Oct. 2010, pp. 301–305.

[19] N. V. Chawla, K.W. Bowyer, L. O. Hall, andW. P. Kegelmeyer, ''SMOTE:

Synthetic minority over-sampling technique,'' J. Artif. Intell. Res., vol. 16,

pp. 321–357, Jun. 2002.

[20] S. U. Habiba, Md. K. Islam, and F. Tasnim, ''A comparative study on

fake job post prediction using different data mining techniques,'' in Proc.

2nd Int. Conf. Robot., Electr. Signal Process. Techn. (ICREST), Dhaka,

Bangladesh, Jan. 2021, pp. 543–546.

[21] G. Othman Alandjani, ''Online fake job advertisement recognition and

classification using machine learning,'' 3C TIC, Cuadernos de Desarrollo

Aplicados a las TIC, vol. 11, no. 1, pp. 251–267, Jun. 2022.

[22] A. Gosain and S. Sardana, ''Handling class imbalance problem using

oversampling techniques: A review,'' in Proc. Int. Conf. Adv. Comput.,

Commun. Informat. (ICACCI), Delhi, India, Sep. 2017, pp. 79–85.

[23] F. Akhbardeh, C. O. Alm, M. Zampieri, and T. Desell, ''Handling extreme

class imbalance in technical logbook datasets,'' in Proc. 59th Annu.

Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang.

Process., 2021, pp. 4034–4045.

[24] J. Ah-Pine and E.-P. Soriano-Morales, ''A study of synthetic oversampling

for Twitter imbalanced sentiment analysis,'' in Proc. Workshop Interact.

Between Data Min. Nat. Lang. Process. (DMNLP), Riva del Garda, Italy,

Sep. 2016, pp. 17–24.

[25] J. David, J. Cui, and F. Rahimi, ''Classification of imbalanced

dataset using BERT embeddings,'' Dalhousie Univ., Halifax,

Canada, Jan. 2020. Accessed: Jan. 2024. [Online]. Available: https://

fatemerhmi.github.io/files/Classification_of_imbalanced_dataset_using_

BERT_embedding.