Ensemble Methods for Early Prediction of Dementia

1st T. Abirami
Department of Information Technology
Kongu Engineering College
Perundurai, Erode, Tamil Nadu
abi.it@kongu.edu

2nd P. Jayadharshini
Department of Artificial Intelligence
Kongu Engineering College
Perundurai, Erode, Tamil Nadu
jayadharshini.ai@kongu.edu

3rd S. Prathiksha
Department of Information Technology
Kongu Engineering College
Perundurai, Erode, Tamil Nadu
prathikshas.21it@kongu.edu

4th S. E. Pranesh Kangeyan,
Department of Information Technology
Kongu Engineering College
Perundurai, Erode, Tamil Nadu
praneshkangeyanse.21it@kongu.edu

5th M. Mohanram,
Department of Information Technology
Kongu Engineering College
Perundurai, Erode, Tamil Nadu
mohanramm.21it@kongu.edu

2024 IEEE 4th International Conference on ICT in Business Industry & Government (ICTBIG) | 979-8-3315-1898-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICTBIG64922.2024.10911083

***Abstract*— Dementia is a progressive cognitive disorder that primarily affects older adults. The patients affected by Dementia experience memory loss, reasoning impairment, and behavioural changes. This condition significantly impacts the daily life and independent activities of the individuals. Although currently available medications cannot reverse or stop the neurodegeneration processes they can temporarily improve or maintain its cognitive functions in the early to moderate phases of the disease. However, early detection of the disease may delay its progression. The paper proposes a systematic approach by applying ensemble methods to predict dementia risk. The algorithms used included XGBoost (XGB), Gradient Boost (GB), Random Forest (RF), Voting Classifier (VC) includes Naïve Bayes (NB) and Support Vector Machine (SVM). The classification algorithms have been compared based on evaluation metrics like accuracy, precision, recall, and f1 score. The model developed achieved an accuracy of 97% and outperformed other methods on the same dataset. The final model enables the early detection of dementia, which is vital for slowing down the neurological decline associated with the disease. This approach is beneficial for underserved populations with limited access to healthcare providers.**

Keywords*— *Dementia, XGBoost, Gradient Boost, Random Forest, Voting Classifier, Naïve Bayes, Support Vector Machine.

I. Introduction

Dementia is a serious condition experienced by millions worldwide as it causes memory and cognitive failures in thought and communication. Dementia is not a disease but a syndrome containing symptoms considered to originate from various types of neurological disorders: the most common is Alzheimer's disease. With the increasing old-age population of the world, this disease is on an exponential rise and seriously challenges health systems, families, and caregivers all over the world. It does not only reach to the individual but also to families and communities as they move along in giving care and support.

According to the World Alzheimer Report 2023, presented by Alzheimer's Disease International, there are already over 58 million people living with dementia in the world. This figure is estimated to rise almost up to 138 million by the year 2050 mainly by population ageing in low-and middle-income countries because it accounts for most cases of dementia. Dementia is also a significant source of economic burden as an annual cost of over $1.3 trillion worldwide may rise to more than $2.0 trillion within the next ten years. The hospitalization, rehabilitation services, and community-based services are part of this cost coupled with substantial informal care given by family members, mostly women while at the same time functioning as caregivers and patients.

The trends observed are appalling; most forms of dementia remain unknown in developing countries, thus denying various people critical care and services, meaning that these issues are addressed through giving remedial education, proper assessment, and quality health-care services-the need to adequately take on this global plague.

All these factors necessitate the development of a model specifically tailored for the detection of dementia as the bedrock for improvement in earlier diagnosis and management of the disease. Detection of dementia will be eased with prompt intervention, which may slow the progression of the disease, improve the quality of life of suffering individuals, and reduce the burden on caregivers as well as healthcare infrastructures. The proposed model, based on advanced machine learning techniques especially ensemble methods, aims at making accurate predictions so that vulnerable populations can be known before it becomes a problem to lead to better treatment decisions and resource allocation.

II. Literature Survey

Recent advances in machine learning have wide ranging implications on the medical field, mainly related to the preliminary diagnosis and forecasting of several diseases. The present study by P.-H. Kuo et al. [1] has aimed at very early diagnosis of dementia. This paper introduces new work in applying machine learning methodology to predict dementia from analysis carried out on brain MRI scans. The research applied various methodologies for enhancement of the model to achieve greater dependability with practical application particularly in such fields where access to healthcare professionals is limited. This system promises rehabilitation strategies to aid improved quality of life among patients by intervention in real time towards recovery. Although the prognosis of dementia is still hard to

979-8-3315-1898-1/24/$31.00 ©2024 IEEE

predict, this approach is a step in the process toward a more effective and feasible screening model.

Tahami Monfared et al. [2] provided comprehensive reviews of the epidemiology and clinical course associated with Alzheimer's disease. In reference to the study, AD falls under one of the highest causes of dementia. This group stated that prevalence and incidence rates will rise with demographic aging. The authors' focus is to emphasize the transition of MCI into Alzheimer's disease (AD) dementia; although 40% to 75% of cases with MCI evolve toward dementia as evidenced by biomarkers, these incorporate amyloid-beta (Ab) deposition and neurodegeneration features. Bianchetti et al. [3] discussed some views around the theme of dementia within the scope of COVID-19. Through this, the research revealed the association of dementia with increased risk of death in Italian patients hospitalized due to COVID-19; such atypical presentations like delirium would demarcate timely medical intervention. Framework for very early detection shall be devised to treat comorbid conditions and predict it on time, support it as well.

D. Panuli et al. [4] reviewed recent trends in cancer diagnosis, with specific attention paid to the methods of machine learning and deep learning, extremely prevalent in the identification and classification of various cancers. Work done by F. M. Talaat et al. [5] presented an Enhanced Dementia Detection and Classification Model, EDCM, along with the help of which this model is feasible to detect at the early stage of dementia in persons whose ages are lesser than 65 using the analysis of segmented images of the brain. Despite that, the study shows some limitations: it mainly focuses on younger populations, which limits its applicability to elderly populations where dementia prevalence is significantly larger; it is hugely based on high-quality MRI data, which are not easily accessible in clinical settings; in addition, the small sample size of young dementia patients raises concerns about the strength and scalability of the model.

Moreover, applications of machine learning also include environmental forecasting, as stated by J. Dong et al. [6] had clearly shown to include XGBoost in short-term rainfall prediction, rectifying biases in statistical weather forecasting and it showed that machine learning exhibits better performance in the different regions. J. Ma et al. [7] have also discussed the initial application of random deep forests tailored for small feature sets, integrating sophisticated entropy metrics to increase the precision and efficacy of the algorithms of machine learning. For neurological diseases, M. D. Benedetto et al. [8] proposed developing deep learning networks to allow the easy identification of first-degree schizophrenia behavioral changes with the help of various datasets, focusing on the vast implications of deep learning in complex multi-aspect data analysis.

In connection with infectious diseases, P. K. Roy and A. Kumar [9] discussed some preliminary approaches toward COVID-19 forecasting by the utilization of transfer learning, which has proved to be the clue to taking timely intervention in epidemics. Joshi et al. [10] designed an ensemble two-tiered machine learning model to predict glioma and classify it under different grades. Ensemble classifiers are used with biomarkers. Later, it shows to be very accurate. The approach for their classification appears promising but the number of the dataset deployed here is rather small; that may potentially be the influence on generalization, hence the finding. Addae et al. [11] discuss the topic of smart technologies - wearable sensor, IoT systems, and machine learning and its application for early detection and management of dementia among elderly patients. These innovations help monitor and predict dementia, thus allowing timely interventions. Their use, however, is limited by accuracy, access, privacy concerns, and high cost, which are specifically significant barriers in a low-resource setting. More integrated and personalized solutions are required for the success of dementia care.

Dhakal et al. [12] paper discusses the prediction of dementia using machine learning with the application of random forest. Important contributors including age, cognitive test scores, and clinical history are emphasized as important predictors; however, the study has some limitations particularly a lack of heterogeneity in the data which limits generalization of the conclusions to broader populations. Second, focusing on structured data overlooks unstructured factors like patient records or lifestyle factors and suggests that model requires further sophistication in the sense that it needs to be more realistic in practical application situations. The results should be validated through larger and more heterogeneous datasets. Ryu et al. [13] proposed a dementia risk prediction model that is based on using the XGBoost algorithm. Along with careful tuning of the hyperparameters, they also managed to obtain derived features. Their work attempted to predict the risk of dementia through brain MRI. Applying the Top-N set of the relevant features improved the framework up to an accuracy of 85.61%. The problem of handling missing data was encountered during the study, which lowered the precision of model prediction.

Another contribution to research on the topic of predicting dementia is that by T. Nguyen et al. [14]. This one involves a paper using machine learning techniques for temporal health record analysis, in which predictions of dementia risk can be made several years before a patient becomes clinically demented. Nyholm et al. [15] followed up on the potential relationship between sleep disturbances and dementia with the help of SNAC data related to populations aged. The paper provides evidence of how machine learning can be applied to enhance early detection capability for dementia through noninvasive monitoring of sleep parameters in real-world clinical settings. However, constraints include that a geographically specific dataset may not apply to diverse populations and emphasizes disturbances of sleep which may otherwise miss out on crucial factors such as genetics and lifestyle.

Such studies aggregate their cumulative findings to point out the need for more comprehensive data in appropriate use. Taking this thought forward, the proposed model utilizes ensemble techniques to make better predictions regarding early dementia onset and takes lessons from previous studies eventually to boost prediction accuracy in this very sensitive area.

III. Methodology

It is the early prediction of dementia with the processing and perfecting of data, which then leads to meaningful strategy formulation. The details within the data are properly maintained and filtered with adjustments to ensure precision. The ensemble methods are then trained and perfected to perfection for optimal outputs. Finally, the finest models are integrated to yield the most efficient prediction. Three measures: precision, accuracy, and recall evaluate the results, which have already confirmed the proposed model to be of excellent performance both in understanding and predicting dementia. Generally, the architecture in the proposed model schematically shows as in Figure 1.

A. Data Collection

The dataset used in this study is the OASIS_longitudinal, downloadable on the Kaggle platform. The dataset contains statistical MRI data of the brain from demented right-handed subjects whose ages range from 60 to 96 years. It consists of a total sample of 150 males and females, and for each of these there were two or more MRI scans taken that were done about one year apart, therefore totaling to 373 scans. The variables used in making the predictions are as outlined in Table 1.

Table I

Data Parameters

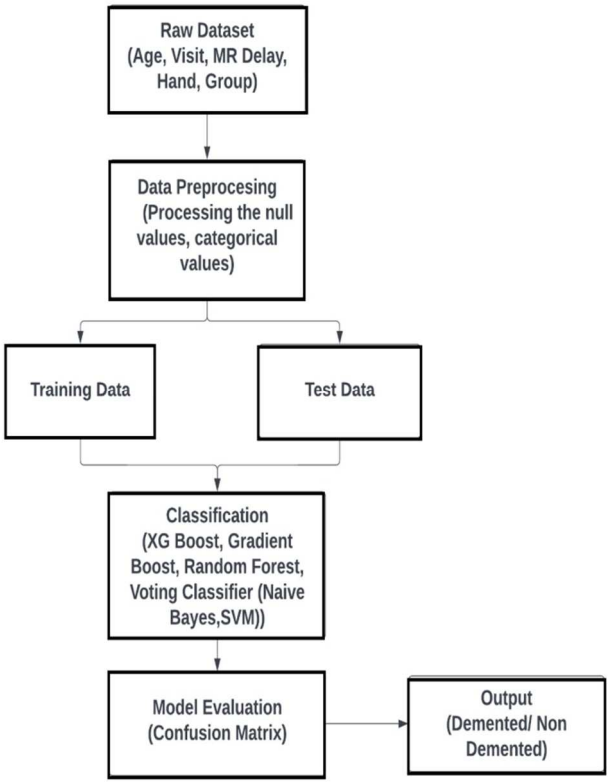
Parameter	Description
Visit	Number of MRI scans
MR Delay	Time Interval between two or more MRI scans
M/F	Sex
Age	Age
EDUC	Years of education
SES	Socioeconomic status: highest status: 1; lowest status: 5
MMSE	Mini-Mental State Examination Score
CDR	Clinical Dementia Rating: 0, normal; 0.5, very mild dementia; 1, mild dementia; 2, Moderately severe dementia; and 3, severe dementia
eTIV	Estimated Total Intracranial volume (from MRI)
nWBV	Normalized Whole Brain Volume (determined using MRI data)
ASF	Atlas Scaling Factor

B. Data Preprocessing

Preprocessing activities were done before the training of models in order to remove variability and unnecessary information. In the data preprocessing, various procedures were carried out to replace missing data and prepare the materials for modeling purposes. In statistical features, missing values have replaced by average of each feature such that the data will not get biased and inconsistent. Categorical features were constructed by using most frequent classes in given feature while taking care about missing values. Interaction term between age and CDR was derived and a multilateral age factor was included, which is Mixed age. Their purpose was to endow the model with a more subtle view of the relationships between variables.

C. Feature Engineering

Feature engineering is the process of extracting only that information from raw data which will be considered by the model. After rectification of the feature set, numerical features go through StandardScaler to standardize them. This method normalized the above parameters to have a mean of 0 and a standard deviation of 1. The standardization during this process is essential since it makes it possible for every statistical variable to be measured on the same scale. Standardization enhances the performance of the anti-sensitivity model as the input data is scaled. OneHotEncoder was used on categorical attributes. This has enabled categorical variables to be transformed into a series of 0s and 1s the model can read and apply within its own method. Then, this approach ensures both statistical and categorical data to well condition for a modeling exercise that has made both model accuracy and efficiency better.



.Fig. 1 General architecture of the proposed model

IV. Module Description

A. Gradient Boost

It is actually a technique for machine learning used to enhance prediction accuracy by combining numerous weak models, which are usually decision trees. This process of

model development begins with an initial training of the model, followed by another model in an effort to correct the mistakes pointed out by the first model. Subsequently, by repeating this process, that is, building the next model with the intention of rectifying the mistakes of the previous one, increasingly accurate predictions are made. The final forecast is thus obtained from summed outputs of all models modified by a learning rate to balance the degree of contribution for each model. Mathematically, if $F_m(x)$ is the prediction at step:

$$F_{m+1}(x) = F_m(x) + \gamma h_m(x) \quad (1)$$

where $h_m(x)$, the next prediction is (x) is the new tree's prediction and γ is the learning rate. This approach allows the model to gradually improve and handle complex patterns in the data.

B. *Random Forest*

In a Random Forest model, individual decision trees form many different subsets of the training data, where every subset has a randomly selected set of features, and so, the variability among the trees is induced, hence combating overfitting. Overfitting is when a model performs well on training data but really poorly on unseen data.

All the constructed trees yield an ultimate result. Random Forest typically makes use of a majority vote in classification and averages the prediction for regression. This approach tends to improve the accuracy because the incorrect predictions of the individual trees primarily cancel each other out.

The mathematical notation to denote the classification predictor of a Random Forest is:

$$\hat{y} = mode(\hat{y}_1, \hat{y}_2, \dots, \hat{y}_B) \quad (2)$$

Here are the predictions from each of the *B* decision *y* trees, and mode is the most common prediction among them.

The average of predictions by all the trees is the final prediction of the regression.

$$\hat{y} = \frac{1}{B}\sum_{i=1}^{B} \hat{y}_i \quad (3)$$

In this equation, $\hat{y}_i$ is the prediction from the i-th tree, and B is the total number of trees in the forest.

C. Voting Classifier

A voting classifier is one of the effective ensemble methods whereby the prediction of a set of machine learning models is combined to give the best result. This is so in the case of a voting classifier, where multiple independent base models or weak learners are trained on the same input data and their individual decisions combined to give a final decision. There are two classes of voting: hard voting and soft voting.

Hence, voting classifiers benefit from the ability to exploit the best models' strengths. In this manner, they avoid overfitting and robustness. The diversity of the models which can have some bias or strength allows the voting classifier to generalize better on unseen data than any individual model. For that reason, voting classifiers are preferred in many applications where the requirements for high accuracy and reliability are critical.

D. XGBoost

XGBoost, or "Extreme Gradient Boosting," is a moderately complex algorithm in machine learning combining the output of many quite simple models into one incredibly powerful predictor, thus enhancing the prediction capability. It supports solving classification and regression problems. The central component of XGBoost remains the objective function, including a loss function and the regularization term that prevents overfitting.

$$Objective = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(\mathrm{f_k}) \quad (4)$$

There depends on the class of problem, the loss function L. For example, in regression problems, it could be the squared error:

$$L(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2 \quad (5)$$

The Equation (5) is the loss function L.

V. Results and Discussion

In the subsequent section, the proposed model will be evaluated in terms of precision and compared with the models in Section III. Moreover, it will also be compared with the other research models that apply the same dataset used in the current study. A comprehensive comparison is presented in Table 2.

A. Comparision of various models

1) Gradient Boost: The Gradient Boosting model brought about 93.62% accuracy-an accumulation of output from different weak learners for clear performance. Precision and recall were both 92.31%, meaning it approached positive instances in a principled manner, failing to miss any. Its F1-Score of 92.31% further strengthens its all-round efficacy. For both classes, the Gradient Boosting model performed uniformly, albeit marginally better scores when it came to negative instances. Figure 2 represents Confusion matrix for the Gradient Boosting model.

TABLE II

COMPARISON OF MODELS

Model	Accuracy	Precision	Recall	F1-Score
GB	93.62	92.31	92.31	92.31
RF	95.74	92.68	97.44	95
VC	96	93.94	96.88	95.38
XGB	97.33	94.12	1.0	96.97

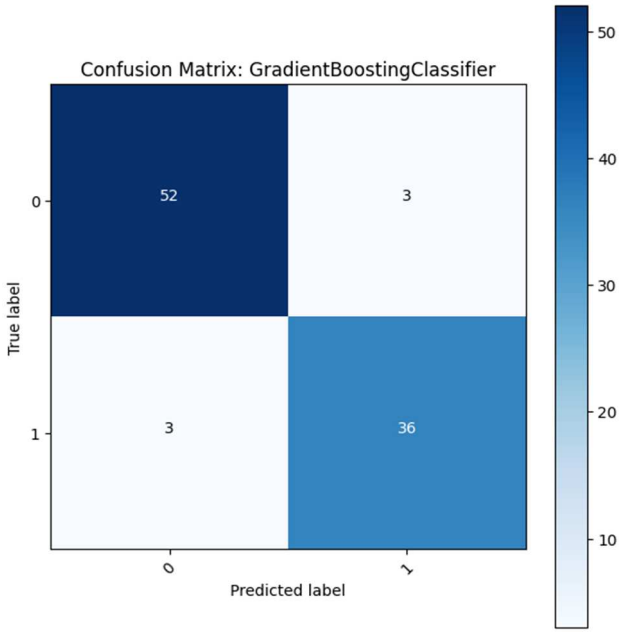


Fig. 2 Confusion matrix of Gradient Boosting

2) Random Forest: The accuracy of the Random Forest model was excellent at 95.74%, which showed this model is robust because it aggregated the prediction result from a set of decision trees. It achieved an accuracy of 92.68%, which, in other words, means that it correctly classified the positive class without a doubt. The recall value is 97.44%, which indicates that the model caught most of the positive instances. The F1-Score reached 95.00%. The classification report indicates that the Random Forest model performed slightly better in terms of recall but compared to precision, so it worked well on both classes. Figure 3 represents the confusion matrix associated with the Random Forest model.

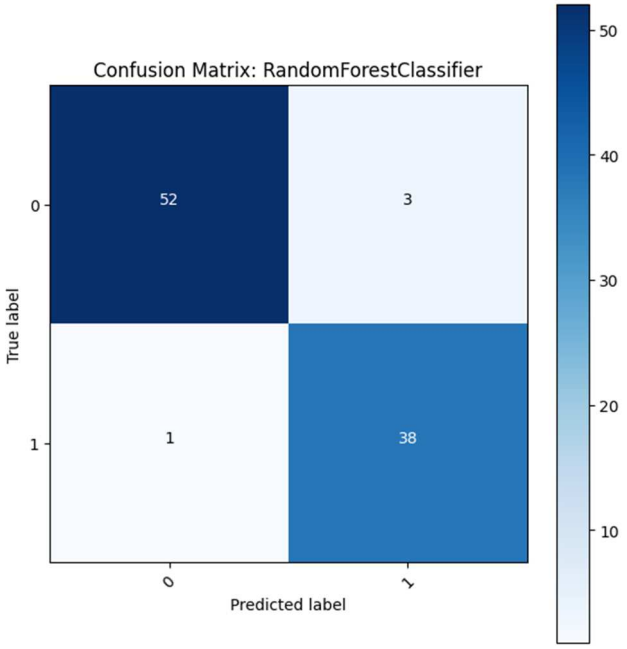


Fig. 3 Confusion matrix of Random Forest

3) Voting classifier: The Voting Classifier uses this approach with two algorithms, SVM and Naive Bayes, which makes the basic predictive power better. This process uses a process called "hard voting," whereby the general votes obtained from each model are used in determining the individual's categorization as 'Demented' or 'Non-Demented.' This model had an accuracy of around 96%, which means that the model predicted correctly in most of its cases. Its accuracy was 93.9%, which made it strong in the discrimination of 'Demented' patients with fewer false positives. Besides this, its good rating was also end with an accuracy of 96.8%, so that it can have a great capability to accurately classify almost all the actual cases of dementia. The Voting Classifier can make better balanced predictions by exploiting both the robustness of SVM and Naive Bayes's simplicity. Therefore, the outcome is a highly effective ensemble model that can become beneficial by capitalizing on the strengths of both the classifiers for the prediction of dementia. Confusion matrix obtained for voting classifier is shown in Figure 4.

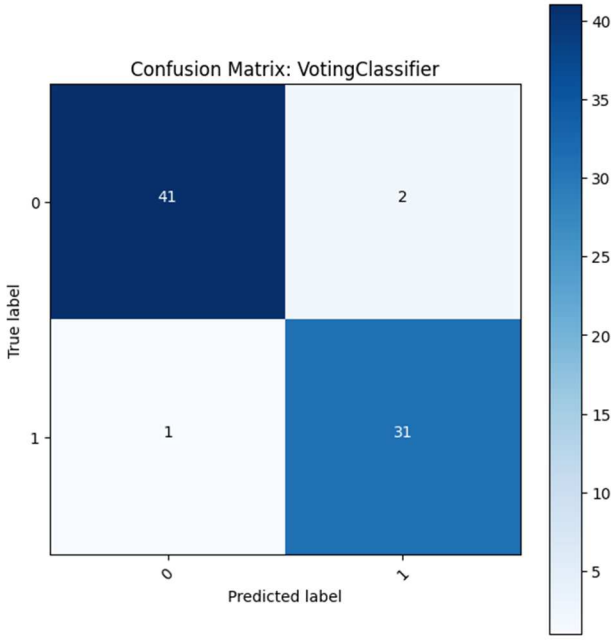


Fig. 4 Confusion matrix of Voting Classifier

4) XGBoost: The XGBoost model also showed the highest accuracy of 97.33% in comparison to the other models, outperforming them. This employs gradient boosting, a concept that enhances its predictiveness by iteratively rectifying its errors, making it particularly effective to be applied on complex datasets. It presents a high precision rate of 94.12%, which indicates the reliability in identifying positive cases in the target class, with an immaculate recall of 100%, implying no missing true positives. The F1-score of 96.97% depicts a robust performance wherein aspects of precision and recall are matched up equally well. The detailed assessment depicts that the model did remarkably good in detecting positive instances and had noticeably high precision and recall values for negative instances as well. The confusion matrix of XGBoost is given in Figure 5.

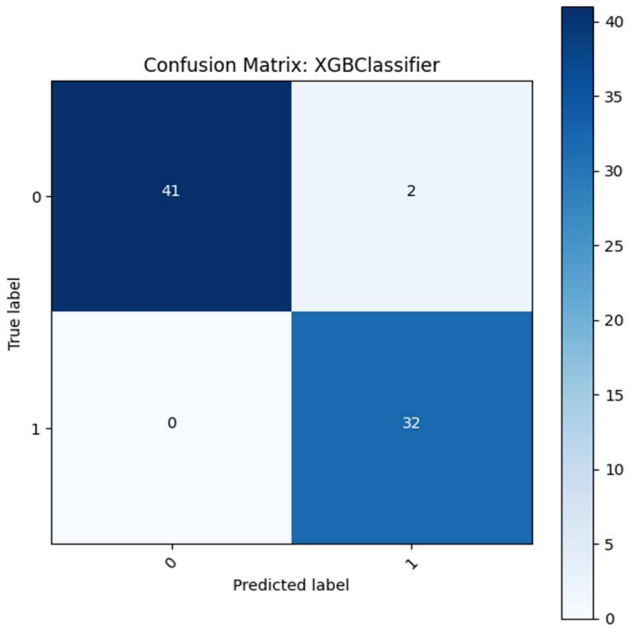


Fig. 5 Confusion matrix of XGBoost

VI. Conclusion

In conclusion, the XGBoost model achieved an impressive figure of 97.33%. Therefore, it had incredible prowess in handling complex data as well as making appropriate predictions with a high F1-Score of 96.97%. The Random Forest model has a good level of accuracy of about 95.74%, making it strong through the ensemble method, though the Gradient Boosting model can assure a fair level of accuracy of 93.62%. The Voting Classifier, which also took advantage of the strength coming from combining SVM and Naive Bayes, was not falling behind either, at 96% accuracy, again striking a balance between precision and recall. Again, the results resonate how well ensemble methods predict dementia, especially when XGBoost is in the top league and it can lay bare the underlying strengths of model combinations to elevate the quality of models in general. Some potential future research directions include, for example, if auxiliary data sources such as neuroimaging or genetic data are also taken into consideration, then the preliminary result will be more accurate information regarding the prediction.

References

[1] P.-H. Kuo, C.-T. Huang, and T.-C. Yao, "Optimized Transfer Learning-Based Dementia Prediction System for Rehabilitation Therapy Planning," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 31, pp. 2047-2059, 2023.

[2] A. A. Tahami Monfared, M. J. Byrnes, L. A. White, and Q. Zhang, "Alzheimer's Disease: Epidemiology and Clinical Progression," Neurology and Therapy, vol. 11, no. 2, pp. 553-569, Mar. 2022.

[3] A. Bianchetti, R. Rozzini, F. Guerini, et al., "Clinical Presentation of COVID-19 Infection in Dementia Patients: A Retrospective Study in an Acute Hospital in Northern Italy," The Journal of Nutrition, Health & Aging, vol. 24, no. 6, pp. 560-562, 2020.

[4] D. Panuli, S. Bhardwaj, and U. Köse, "Recent advancement in cancer diagnosis using machine learning and deep learning techniques: A comprehensive review," Comput. Biol. Med., vol. 146, Jul. 2022.

[5] F. M. Talaat and M. R. Ibraheem, "Dementia diagnosis in young adults: a machine learning and optimization approach," Neural Computing and Applications, vol. XX, pp. 1-10, 2024.

[6] J. Dong, W. Zeng, L. Wu, J. Huang, T. Gaiser, and A. K. Srivastava, "Enhancing short-term forecasting of daily precipitation using numerical weather prediction bias correcting with XGBoost in different regions of China," Eng. Appl. Artif. Intell., vol. 117, Jan. 2023.

[7] J. Ma, Q. Pan, and Y. Guo, "Depth-first random forests with improved grassberger entropy for small object detection," Eng. Appl. Artif. Intell., vol. 114, Sep. 2022.

[8] M. D. Benedetto et al., "Deep networks for identification of behavioral variant frontotemporal dementia from multiple acquisition sources," Comput. Biol. Med., vol. 148, Sep. 2022.

[9] P. K. Roy and A. Kumar, "Early prediction of COVID-19 using ensemble of transfer learning," Comput. Electr. Eng., vol. 101, Jul. 2022.

[10] R. C. Joshi, R. Mishra, P. Gandhi, V. K. Pathak, R. Burget, and M. K. Dutta, "Ensemble based machine learning approach for prediction of glioma and multi-grade classification," Comput. Biol. Med., vol. 137, Oct. 2021.

[11] S. Addae, J. Kim, A. Smith, P. Rajana, and M. Kang, "Smart solutions for detecting, predicting, monitoring, and managing dementia in the elderly: A survey," IEEE Access, vol. 12, pp. 1-10, 2024.

[12] S. Dhakal, S. Azam, K. M. Hasib, A. Karim, M. Jonkman, and A. S. M. F. Al Haque, "Dementia prediction using machine learning," Procedia Computer Science, vol. 219, pp. 1297-1308, 2023.

[13] S.-E. Ryu, D.-H. Shin, and K. Chung, "Prediction model of dementia risk based on XGBoost using derived variable extraction and hyper-parameter optimization," IEEE Access, vol. 8, pp. 177708-177717, Sep. 2020.

[14] T. Nguyen, H. Tran, and S. Kim, "Predicting Dementia Risk with Machine Learning and Longitudinal Health Data," IEEE Journal of Biomedical and Health Informatics, vol. 26, no. 5, pp. 1942-1952, 2022.

[15] J. Nyholm, A. N. Ghazi, S. N. Ghazi, and J. S. Berglund, "Prediction of dementia based on older adults' sleep disturbances using machine learning," Computers in Biology and Medicine, vol. 171, pp. 1-10, 2024.