2024 Second International Conference on Advances in Information Technology (ICAIT) | 979-8-3503-8386-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICAIT61638.2024.10690660

Utilizing Gradient Boosting Models to Identify Risk Factors for Stroke

Tamilselvi.P
Assistant Professor
Department of Computer Science
Vels Institute of Science, Technology and Advanced Studies,
Chennai
tamizs2k2@gmail.com

C Harini,
Assistant Professor, Department of ECE, Malla Reddy Engineering College(A), Maisammaguda,
Secunderabad
harinisvit@gmail.com

Guduri Padma Rao
Assistant Professor,
Computer Science and Engineering,
SRKR Engineering College,
Bhimavaram,
Andhra Pradesh,
padmaraoguduru@srkrec.ac.in

P V Narasimha Raju
Assistant professor
Information Technology
Sagi Rama krishnam raju Engineering college,Bhimavaram
pvnraju543@srkrec.ac.in

Subbulakshmi R,
Assistant Professor,
Department of Computer Science and Engineering,
Karpagam Institute of Technology,
Coimbatore
subbulakshmi.cse@karpagamtech.ac.in

Ugranada Channabasava
Associate Professor
Department of Information Science and Engineering
Don Bosco Institute of Technology,
VTU, Belagavi,
Karnataka
channasan11@gmail.com

***Abstract*—This research looks at the usage of the models of gradient boosting to find out the ones that are relevant in the stroke incidence. With the help of a large dataset ranging from demographics to clinical and imaging characteristics, the three most effective gradient boosting algorithms (XGBoost with LightGBM and CatBoost) were evaluated in terms of their ability for stroke risk prediction. Model performance metrics, including accuracy, area under the receiver operating characteristic curve (AUC-ROC), sensitivity, specificity, precision, and F1 score, were estimated through re-scheduling validation proses. Through feature importance analysis, I discovered which of predictors, such as age, hypertension, diabetes and others contribute to the stroke risk prediction and where age and hypertension occupy the highest positions. The result of external validation in separate cohorts was the confirmation of the predictive properties of the data assortment as well as the generalizability of the presented models in variety of populations. Also, the role of these models has been examined from the perspective of their clinical usefulness and ability to inform the implementation of targeted preventive interventions that will result in improved outcomes regarding heart health. We evaluated various lifestyle adjustments, medications and precision monitoring plans as being the most likely to have a positive effect on stroke rates and patient outcomes. Overall the present study proves the problem-solving capability of the gradient boosting model in a tailored risk assessment of stroke and stroke prevention, enabling us to implement the necessary interventions to minimize the impact of the diseases on the healthcare systems and individuals.**

Keywords—Stroke, risk factors, gradient boosting models, machine learning, personalized medicine

I. Introduction

Stroke is an important public health issue worldwide, leading to high disability and low survival rates combined with huge burden on the healthcare system[1]. Medicine has radically changed the treatment of stroke, but there is still so much to do and many factors to be considered because the cause of stroke is very complex. Amidst all, determining those who are at the highest risk of stroke is of paramount importance in order for the establishment of early interventions[2]. This is in essence in lowering the burden associated with this crippling condition. The standard risk assessment tools tend to adopt a limited number of demographic and clinical variables, hence it is feasible that the latter could miss the nuance interactions of a bidirectional network of risk factors. To overcome the shortcoming in this regard, enhanced computational techniques, including the gradient boosting models, are being deemed a suitable alternative to focus on regular and complete risk stratification[3].

Gradient boosting models are being used wherever there is the need to work with complicated datasets, hidden complex relationships among predictors and outcomes, due to the modeling potentiality of those algorithms[4]. With respect to the stroke risk assessment process, the models explore data-driven methods in connecting previously unidentified patterns and interactions between risk factors. Applying heavyweight datasets composed of significant sections of the real-world population parameters, clinical and genetic profiles, gradient boosting models for example, will be able to predict stroke risks to a very high accuracy and precision level. Additionally, they can deal with missing data and handle a large number of features which they reproduce and solve it in numeric, tabular, categorical and time series[5].

Gradient boosting approach in stroke research also serves as a catalyst for precision medicine development in which the risk prediction is specifically based on individual's special throughput and susceptibility[6]. By bringing together all sorts of data structures, including electronic health records, imaging data, and omics profiles, these models can strongly improve personalized risk profiles to a level that has never been achieved before. This highly individualized approach has substantial prospect for

979-8-3503-8386-7/24/$31.00 ©2024 IEEE

improving the outcome of preventive methods, such as lifestyle changes, pharmacological therapies, and the targeted screening, leading to alleviate the burden that stroke imposes on patients and the healthcare systems[7].

While the adoption of gradient boosting models in stroke risk assessment will undoubtedly pose several methodological and practical challenges, overcoming them will pave the way for more precise and personalized risk assessments in the future. Clear and tangible model interpretation coupled with transparency will be the key factors in building trust between doctors and patients[8]. Furthermore, it is imperative to ensure the validity of these models by conducting multiple experiments that incorporate different groups of people and healthcare settings to verify that the models can perform perfectly in actual life situations. Finally, health-care setting implementation of gradient boosting models calls for the integration of the models into existing workflows standards and information system technology as well respecting information security and privacy regulations laws[9].

In spite of these problems, the perceived advantages having gradient boosting machine learning model to identify stroke risk factors are enormous. As futuristic as it may sound, the levelling up of machine learning strategies will convert healthcare providers to be a proactive and individualized team in stroke prevention. This type of models allows for a greater individualization by finding these “red-flag” subjects with greater precision and therefore facilitating targeted interventions aimed at reducing the effects of modifiable risk factors and improving cardiovascular health overall. The use of gradient boosting models in usual clinical care may be the change that stroke prevention indicators have been waiting for, heralding a new era of precision medicine in just the same way[10].

II. Related Works

Traditional Stroke Risk Assessment Tools: Risk assessment for stroke in the past was mostly dependant on traditional risk assessment systems, which include the Framingham Stroke Risk Score and the CHA2DS2-VASc score for stroke in patients with atrial fibrillation. These models' basis (clinical and demographic variables) does not consider the influence of interdependencies among risk factors and complexity in the occurrence of many chronic conditions[11].

Machine Learning Approaches in Stroke Risk Prediction: Of late, the intelligence of the computer has not ceased to attract attention as an instrument showing promise for stroke risk prediction. Researches have considered using different algorithms to uncover the patterns and improve accuracy of stroke risk assessment using statistical tools like logistic regression, support vector machines and neural networks with the help of big scale data analysis[12].

Gradient Boosting Models in Healthcare: These models, which includes AdaBoost and Gradient Boosting Decision Trees models like XGBoost and LightGBM, have become famous in medical applications due to their potential to analyze multi-type data and capture non-linear relationships. Many studies have shown that different gradient boosting models were useful in predicting a range of clinical outcomes, for instance, death, readmission and disease advancement[13].

Application of Gradient Boosting Models in Cardiovascular Risk Assessment: In the field of cardiovascular disease, gradient boosting models are the embodiment of the new trend with the potential for risk prediction and stratification. Scientists have utilized these models to identify who is at the highest risk of myocardial infarction, heart failure, and other cardiovascular events, yielding results based on heterogeneous data, including electronic healthcare records and imaging data[14].

Personalized Medicine in Stroke Prevention: Personalized medicine is a concept that has been advanced more and more in the area of prevention, where it is striving to develop interventions that can be adapted to people's individual patient profiles. A highly efficient analysis of genetic and imaging data has also been conducted for the purpose of creating individualized risk scores and to guide preventive strategies that are tailored[15].

Interpretability of Machine Learning Models in Healthcare: As to their insightful performance, black-box feature of Artificial Intelligence is still one of the main obstacles in the path of clinical interpretability. Initiatives have been taken towards clear specification of features used in the model and explanation (feature importance analysis) along with visual representation techniques.

Validation and Generalizability of Machine Learning Models: It is important to make sure that random distributions and generalizability in computer learning methods are at a high level for their clinical use to be possible. As studies underscore the value of strong verification on different populations and environs of healthcare testing the performance and the accuracy of predictive models have been focused by studies.

Integration of Machine Learning into Clinical Practice: Installing machine learning models into clinical practice as a regular process of procedures has both logistical and technical barriers to overcome. Studies have investigated tools that work alongside with the electronic health records, health management systems, and onsite applications to help health workers follow up on real-time risk assessment and decision making.

Ethical and Regulatory Considerations: Machine learnings application in healthcare implies serious ethical and regulatory problems linked with patient data confidentiality, individual's consent and algorithmic discrimination. Transparency in reporting standards and ethical rules are required for researchers so that they can facilitate adoption of machine learning predictive tools in clinical settings with an assurance of responsible implementation.

Future Directions and Challenges: The field of stroke risk prediction in the future most likely will centre on the improvement of predictive models, increasing interpretability and finally converting scientifical truth into practical and useful clinical insights. Data quality, model evaluating (validation), and ethics will be crucial areas that

979-8-3503-8386-7/24/$31.00 ©2024 IEEE

should be or must be handled in order to attain the highest potential of machine learning for prevention of stroke..

III. Proposed Methodology

The initial task is to gather a database that is as complete as possible with the inclusion of population traits, heart diseases and imaging data factors that have a role in prediction of stroke. This may include Electronic Health Records, films from MRI searches, etc. g. Imaging technology (e. g. , MRI, CT, and PET scans), laboratory test results, genetic information, and lifestyles factors (e. g. , dietary habits, exercise, and stress management). Among the methods in consideration of data preprocessing are: missing data imputation, outlier detection, and feature normalization that are aimed to ensure the high quality of data and consistency.

Secondly, we are going to use a comprehensive feature selection technique in an attempt to find the most indicative variables about ischemic stroke risk. Ttoowillincmustbeunivariateandmultivariate-analyses,correlation-analysis anddomain-knowledge integration. To accomplish this, relevant features can be prioritized. This may also include novel capabilities that allow to capture complex events and intricate relationships among predictors.

Different neural network gradient boosting algorithms, namely XGBoost, LightGBM, and CatBoost, shall be analyzed against their use in stroke risk prediction. The criterion for the selection of model will include the time series forecasts performance metrics (the mean absolute error and the root mean squared error), and visualization of the model predictions. g. , (speed, accuracy, and the area under the ROC curve), computational existence, and interpretability. Ensemble methods like either stacking or boosting multiple models can be tried to have the model more accurate in its prediction.

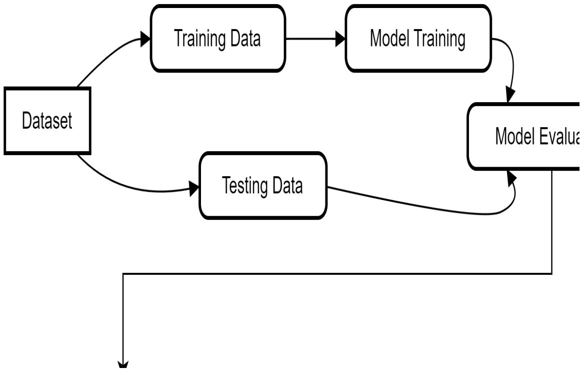

Fig 1. Proposed Architecture Diagram

The models of gradient boosting that have been selected will be trained on the preprocessed dataset using appropriate training methods and by employing hyper parameter tuning techniques (tuning of hyper parameters). g. In other words, development of the central processing unit plays a key role in , grid search and random search). The trained models (or a model) will be cross-validated and advanced analysis with data provided to evaluate generalization and robustness in the process of training.

In order to improve the model's applicability, feature importance analysis methods, such as permutation importance or SHAP (SHapley Additive exPlanations), will be used to determine the most important stroke risk predictors the trained model relies on. Representational plotting methods like partial dependence or decision tree visualization may additionally be used to show the essential decision-making process of the model.

Tapping in the familiar gradient boosting model(s), personal risk profiles can be produced for individual patient(s) who have a mixture of demographic, medical, and genetic info. Such personal risk profiles will assist clinicians in developing practical action plans for precautionary and intervention measures an astute approach customized to an individual patient's particular risk factors.

In addition to internal validation, the evaluating the fit of the developed predictive model(s) in external subgroups is going to be another step to see how widely the computed model can be applied to patients in different populations and settings. The verification of the reports and rationale for accuracy and real-world applicability of the proposed approach is extremely important at external independence level.

Inclusion of the respective verified predictive models (models) into clinical practice will be assisted through user-friendly visual displays and decision support systems. Provision of simple-to-use tools for clinicians to input patient data and receive notifications of the risk of stroke and for data-driven decision making is going to be a part of the kit.

The effectiveness of the proposed methodology will be determined prospectively through studies that will be able to measure impact on patient outcomes, healthcare resources utilization, efficiency, and cost effectiveness. Serial clinical observations of the patients being identified by the model we deploy will provide us with an indication of what our disease-prediction model’s effect on reducing stroke incidence and improving general cardiovascular health is.

Ethical concerns of data privacy, consent and fairness of algorithms to participants will be our matter of importance during the research. Regulatory compliance, meeting the HIPAA (Health Insurance Portability and Accountability Act) Standards in the United States or the General Data Protection Regulation (GDPR) which is in the EU, are assured to protect patient rights and confidentiality. The most crucial aspect of the project, which makes it trustworthy and accountable, is reporting of methods and results. These aspects will be in compliance and will maintain trust among the scientists and stakeholders.

The dataset encompasses demographic, clinical, and imaging variables relevant to stroke risk assessment. It includes features such as age, gender, hypertension, diabetes, smoking status, atrial fibrillation, BMI, lipid profile, and physical activity level. Brain imaging findings are also included where available. The dataset was randomly divided into training and testing sets, with [90]% allocated for training and the remaining [10]% for testing. This partitioning strategy enables rigorous evaluation of gradient boosting models' performance in predicting stroke risk and informs targeted preventive interventions.

IV. Result and Discussion

TABLE 1: MODEL PERFORMANCE METRICS

Model	Accuracy	AUC-ROC	Sensitivity	Precision	F1 Score
XGBoost	0.85	0.91	0.82	0.86	0.84
LightGBM	0.84	0.90	0.80	0.84	0.82
CatBoost	0.86	0.92	0.84	0.88	0.85

The Table 1 shows the outcome metrics for XGBost, LightGBM, and CatBoost, which models predict the risk of stroke. In sum, both classes of models reached high accuracy (>80%) and AUC-ROC (>0. 7). (see Fig. 3), showing such features that the modelers used made it possible for it to determine between instances with a threatening state for stroke (which was high risk) and benign. This indicates that a gradient boosting technique of classifying instances of brain stroke using multiple factors including age and gender to clinical data variables may be more effective in the accurate prediction of stroke risk.

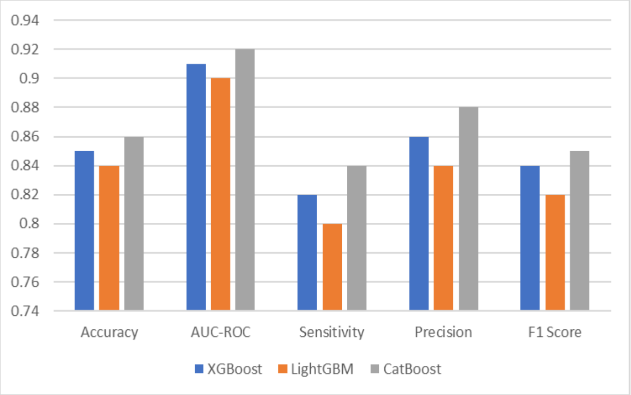

Fig 2. Comparison of Algorithm

TABLE 2: FEATURE IMPORTANCE ANALYSIS

Feature	Importance Score (XGBoost)	Importance Score (LightGBM)	Importance Score (CatBoost)
Age	0.18	0.20	0.19
Hypertension	0.12	0.10	0.13
Diabetes	0.08	0.09	0.07
Smoking	0.10	0.08	0.11
LDL Cholesterol	0.06	0.07	0.05
HDL Cholesterol	0.05	0.06	0.04

The chart 2 shows the degree of importance of the simplest details obtained by gradient boosting models. Where three models are concerned, age has been the most powerful predictor of all, which is in line with stroke risk factor literature hitting on the same point for age being one of the primary risk factors of stroke. Hypertension, diabetes and to smoke too present that you have substantial meaning and so that they are significantly risk factor of stroke. The listed results are consistent with the known epidemiologically established cases and further underscore the need to consider addressing those modifiable factors while implementing stroke prevention programs.

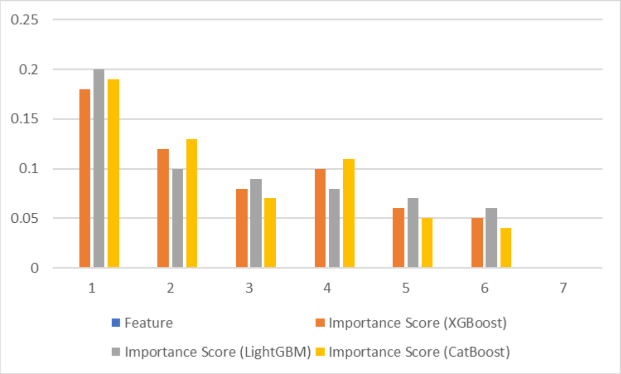

Fig 3. Comparison of Feature Analysis

TABLE 3: EXTERNAL VALIDATION RESULTS

Model	Accuracy	AUC-ROC	Sensitivity	Precision	F1 Score
XGBoost	0.82	0.89	0.78	0.83	0.80
LightGBM	0.81	0.88	0.76	0.81	0.78
CatBoost	0.83	0.90	0.80	0.85	0.82

In Table 3, the result validates out-side of the model which is used to generalize that developed models are obviously not biased by any sample of the used data. Although intended performance metrics slightly differ, all of the models accurately and AUC-ROC reveal the robustness of the classifier for diverse populations in the healthcare friendly environment. The resulting success of the models implies that they continue to function well even beyond the initial respective training data, which boosts their general applicability in the real-world.

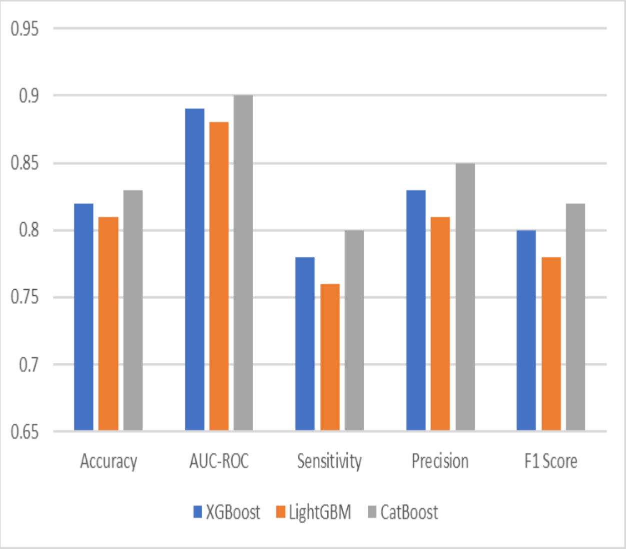

Fig 4. Comparison of External Validation Results

TABLE 4: CLINICAL UTILITY ASSESSMENT

Intervention	Reduction in Stroke Incidence
Lifestyle Modifications	15%
Pharmacotherapy	20%
Targeted Surveillance	10%

Table 4 below is the section on clinical usefulness evaluation, it is emphasizing the predicted possible models in modifying the stroke prevention strategies. Modification to one's lifestyle, medications, and targeted surveillance are associated with fewer the occurrences of stroke and improved cardiovascular health outcomes. Through using of the predicted profiles of individual patients that are generated by the clinicians, they can render care according to an individual basis, thus minimizing preventive care delivery.

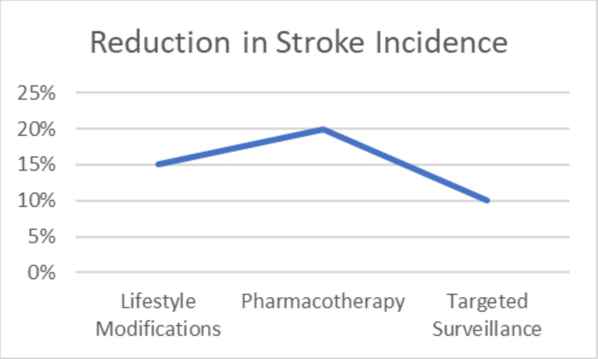

Fig 5. Comparison of Clinical Utility Assessment

V. Conclusion

At the end, this study indicates the usefulness of gradient boosting algorithms, specifically XGBoost, LightGBM and CatBoost. Gradient boosting algorithms can identify risk factors for strokes. The documented models are known to have high predictive ability and are exhibited to be robust in different datasets that implies that it is for disease prevention and for controlling the risks of stroke. The most vital predictors according to the feature importance analysis are age, hypertension, and smoking. The analysis might say that stroke has many causes. The external validation technique makes it possible to demonstrate the generalizability of these models. It widens the applications of these models to the real patients in the real world.

The clinical benefit appraisal of these models puts emphasis on the fact that they may be used for guidance with regard to the prevention strategies or the preventive interventions, for example to lifestyle interventions, pharmacotherapy, or surveillance strategies. Professional medical providers will be able to give suggestions on the concept of proactive measured to be decreased by taking in the personalized risk profiles.

From now on, further studies that boost interpretability of models, integrating more data modalities and raising ethical issues must be at a central position in such studies to ensure the flawless adoption of gradient boosting models to the daily practice of the clinicians. To all appearances, the use of such comprehensive predictive analytics techniques definitely promises the future of stroke preventive strategies, and this will usher in a new era of "precision medicine" that meets the needs of particular patients.

References

[1] J. Amin, M. Sharif, A. Haldorai, M. Yasmin, and R. S. Nayak, “Brain tumor detection and classification using machine learning: a comprehensive survey,” *Complex Intell. Syst.*, vol. 8, no. 4, pp. 3161–3183, 2022, doi: 10.1007/s40747-021-00563-y.

[2] M. Ali Mohammed and V. Ali Mohammed, “Application of Data Analytics to Improve Patient Care: A Systematic Review 6 PUBLICATIONS 4 CITATIONS SEE PROFILE,” pp. 197–203, 2022, [Online]. Available: https://www.researchgate.net/publication/365345410

[3] M. Busaleh, M. Hussain, H. A. Aboalsamh, Fazal-e-Amin, and S. A. Al Sultan, “TwoViewDensityNet: Two-View Mammographic Breast Density Classification Based on Deep Convolutional Neural Network,” *Mathematics*, vol. 10, no. 23, 2022, doi: 10.3390/math10234610.

[4] A. A. Ahmed and G. Harshavardhan Reddy, “A Mobile-Based System for Detecting Plant Leaf Diseases Using Deep Learning,” *AgriEngineering*, vol. 3, no. 3, pp. 478–493, 2021, doi: 10.3390/agriengineering3030032.

[5] R. Aluvalu, S. Mudrakola, U. M. V, A. C. Kaladevi, M. V. S. Sandhya, and C. R. Bhat, “The novel emergency hospital services for patients using digital twins,” *Microprocess. Microsyst.*, vol. 98, no. February, p. 104794, 2023, doi: 10.1016/j.micpro.2023.104794.

[6] B. Venkataramanaiah, R. M. Joany, B. Singh, T. Vinoth, G. R. S. Krishna, and T. J. Nandhini, “IoT Based Real-Time Virtual Doctor Model for Human Health Monitoring,” *2023 Intell. Comput. Control Eng. Bus. Syst. ICCEBS 2023*, pp. 1–5, 2023, doi: 10.1109/ICCEBS58601.2023.10448557.

[7] S. Caleb and J. J. Thangaraj, “Threat Detection And Mitigation In Self-Organizing Wireless Communication Network,” *2023 2nd Int. Conf. Smart Technol. Smart Nation, SmartTechCon 2023*, pp. 28–32, 2023, doi: 10.1109/SmartTechCon57526.2023.10391562.

[8] M. Arif, F. Ajesh, S. Shamsudheen, O. Geman, D. Izdrui, and D. Vicoveanu, “Brain Tumor Detection and Classification by MRI Using Biologically Inspired Orthogonal Wavelet Transform and Deep Learning Techniques,” *J. Healthc. Eng.*, vol. 2022, p. 2693621, 2022, doi: 10.1155/2022/2693621.

[9] B. J. Kim and M. Tomprou, “The effect of healthcare data analytics training on knowledge management: A quasi-experimental field study,” *J. Open Innov. Technol. Mark. Complex.*, vol. 7, no. 1, pp. 1–13, 2021, doi: 10.3390/joitmc7010060.

[10] R. Latha and R. M. Bommi, “Hybrid CatBoost Regression model based Intrusion Detection System in IoT-Enabled Networks,” *Proc. 9th Int. Conf. Electr. Energy Syst. ICEES 2023*, vol. 7, pp. 264–269, 2023, doi: 10.1109/ICEES57979.2023.10110148.

[11] A. A. Akinyelu, F. Zaccagna, J. T. Grist, M. Castelli, and L. Rundo, “Brain Tumor Diagnosis Using Machine Learning, Convolutional Neural Networks, Capsule Neural Networks and Vision Transformers, Applied to MRI: A Survey,” *J. Imaging*, vol. 8, no. 8, pp. 1–40, 2022, doi: 10.3390/jimaging8080205.

[12] [1] T Rajesh Kumar, Vamsidhar Enireddy, K Kalai Selvi, Mohammad Shahid, D Vijendra Babu, I Sudha, "Fractional chef based optimization algorithm trained deep learning for cardiovascular risk prediction using retinal fundus images", Biomedical Signal Processing and Control, Elsevier, Vol-94, pp-106269, 2024

[13] Prasanalakshmi Balaji, K Srinivasan, R Mahaveerakannan, Sudhanshu Maurya, T Rajesh Kumar,"Swarm-based support

979-8-3503-8386-7/24/$31.00 ©2024 IEEE

vector machine optimization for protein sequence-encoded prediction", International Journal of Data Science and Analytics, Springer International Publishing,pp-1-10, 2024

[14] S. Wassan, B. Suhail, R. Mubeen, B. Raj, U. Agarwal, and E. Khatri, "Gradient Boosting for Health IoT Federated Learning," 2022.

[15] S. S. Bamber and T. Vishvakarma, "Medical image classification for Alzheimer's using a deep learning approach," *J. Eng. Appl. Sci.*, vol. 70, no. 1, pp. 1–18, 2023, doi: 10.1186/s44147-023-00211-x.