An explainable and exploratory research utilising a web application for early intervention and machine learning-based stroke prediction.

Prof. D. V. Varaprasad, M.Tech, (Ph.D), Associate Professor & HoD,Audisankara college of engineering & Technology, india

Dr.A.M.Mahaboob Basha, Associate Professor, Department of CSE,Audisankara college of engineering &Technology ,india

GHV Prasad Babu,Department of CSE,Audisankara college of engineering & Technology, india

Abstract: Early medical action and better patient outcomes depend on early prediction of stroke. Extensive ensemble methodology incorporating Categorical Boosting and Stacking Classifier is presented in this work building on current machine learning-based techniques, obtaining an outstanding 99% accuracy. Through efficient processing of many datasets, the proposed approach improves model resilience and guarantees interpretability using SHAP-based explanations. A Flask-based web application with user authentication is designed to guarantee safe access to stroke prediction insights and hence enable practical implementation. Early stroke detection in healthcare systems would benefit much from this extension's greatly enhanced prediction accuracy, security, and usability.

Index terms *-—Stroke Prediction, Machine Learning, Ensemble Methods, Categorical Boosting, Stacking Classifier, Explainable AI, SHAP, SMOTE, Flask Web Application, User Authentication*

1. INTRODUCTION

I. INTRUCTION

The life-threatening medical disorder known as stroke results from a disturbance of blood flow to the brain causing extreme neurological problems. Still one of the primary causes of disability and mortality globally, it calls for early identification and quick response. Conventional diagnostic techniques mostly depend on clinical evaluations and medical imaging, which could be time-consuming and unavailable in environments with limited resources. Automated stroke prediction models have surfaced as a potential option to support early diagnosis and enhance patient outcomes with developments in artificial intelligence (AI) and machine learning (ML).

Recent work has concentrated on creating machine learning models utilising methods such Decision Trees (DT), Stochastic Gradient Descent (SGD), K-Nearest Neighbours (KNN), Support Vector Machines (SVM), and XGBoost. These models' generalising capacity is affected, nevertheless, by unbalanced datasets, lack of interpretability, and reliance on particular hospital data. This work uses an ensemble-based method combining Categorical Boosting and Stacking Classifier to solve these problems by raising prediction accuracy. These cutting-edge methods improve model resilience, therefore enabling better handling of many datasets.

Moreover, a major feature of applications driven by artificial intelligence in healthcare is interpretability. The work uses SHAP (Shapley Additive Explanations) to give model decision-making transparency, therefore guiding medical practitioners in their understanding of the elements affecting stroke predictions. Furthermore designed to provide safe and simple access to automated stroke predictions is a Flask-based web application with user authentication. This study introduces a complete framework that improves stroke prediction accuracy, security, and usability by merging sophisticated ensemble methods, data balancing approaches, and explainable artificial intelligence, therefore opening the path for early medical intervention and better healthcare outcomes.

2. LITERATURE SURVEY

a) Plasmatic retinol-binding protein 4 and glial fibrillary acidic protein as biomarkers to differentiate ischemic stroke and intracerebral hemorrhage:

https://pubmed.ncbi.nlm.nih.gov/26526443/

Quick distinction between acute ischaemia stroke and intracerebral haemorrhage (ICH) can help to improve therapy and outcomes. Our aim was to identify fresh plasma biomarkers able to distinguish between many types of stroke and apply them in combination with present markers for this therapy indication. Eleven of the plasma samples tested using 177 antibodies had varying degrees of chemokines, growth factors, and angiogenic factors across the many kinds of stroke ($p < 0.05$), compared to 36 patients with ischaemic stroke and 10 patients with ischaemic cardiomyopathy. Out of five proteins studied in sixteen patients with ischaemia stroke and sixteen patients with ICH, replicated results were obtained for pigment epithelial-derived factor, apolipoprotein B100, and RPB4 (p Examined in 38 samples from ischaemic stroke and 28 samples from ICH were these proteins, GFAP, and the receptor for advanced glycation end product. RBP4 >61 µg/mL and GFAP <0.07 ng/mL displayed finally 100% subtype specificity. Multivariate logistic regression analysis revealed that GFAP <0.07 ng/mL and RBP4 >48.75 µg/mL were also independent predictors of stroke type, therefore enhancing discrimination by 29% ($p < 0.0001$; ORadj: 6.09, $p = 0.02$). These two markers might help to distinguish ICH from ischaemic stroke. Fast differentiation between ischaemia stroke and intracerebral haemorrhage is absolutely essential for suitable therapy. We identified and validated as plasmatic biomarkers for subtyping hyperacute stroke RBP4 and circulating GFAP. By hastening the classification of stroke subtypes, the use of these and other biomarkers might help to improve patient outcomes.

b) Stroke Risk Prediction with Machine Learning Techniques

https://pmc.ncbi.nlm.nih.gov/articles/PMC9268898/

A stroke results from a sudden cut off of blood flow in the brain. Depending on the location affected, loss of circulation causes brain cells to die and functions to be compromised. Early symptom recognition helps one to predict a stroke and enhance their health. This study generates and tests many ML models to offer a consistent approach for estimating the stroke probability. Among the measures of performance of the stacking technique proposed in this work are accuracy, precision, recall, AUC, and F-measure. Stackering classification was demonstrated to be the most successful technique in the research with a 98.9% AUC, 98% F-measure, precision, recall, and accuracy.

c) Segnet: A deep convolutional encoder-decoder architecture for image segmentation

https://ieeexplore.ieee.org/document/7803544

Using deep fully convolutional neural networks for semantic pixel-wise segmentation, the creative and pragmatic SegNet architecture The 13 convolutional layers of VGG16 have topological similarity to an encoder network [1]. The decoder network generates full-resolution input feature maps from encoder feature maps when it comes time to individually identify pixels. SegNet's decoder upsamples its lower resolution input feature map using an unusual technique. The decoder does non-linear upsampling using the pooling indices following the encoder's max-pooling phase. Upsampling training is thus disabled. Sparse upsampled maps are convolved using trainable filters to create dense feature maps. Among the notable substitutes for this architecture are DeconvNet [4], FCN [2], and DeepLab-LargeFOV [3]. This comparison shows how closely accuracy and memory performance in segmentation correlate. Scene comprehension drove SegNet among other uses. Regarding inference, memory and computing efficiency are its main design objectives. Stochastic gradient descent allows one to train it end-to-end with a far fewer number of trainable parameters than in previous designs. Used as controllers for SegNet and other systems, SUN RGB-D indoor scene segmentation and road scene benchmarking These quantitative assessments show SegNet performs better in terms of inference time and memory efficiency than rival designs. See a live demo of our Caffe SegNet implementation at http://mi.eng.cam.

d) Recent progress in semantic image segmentation:

https://arxiv.org/abs/1809.10198

Semantic picture segmentation is a major application of computer vision and image processing used in both medical and intelligent transportation. Many times, academics test algorithm performance using many benchmark datasets. Semantic segmentation research has been active for some considerable period. Since Deep Neural Networks (DNNs) have emerged, segmentation has evolved. In this work, we classify semantic picture segmentation methods as either conventional or contemporary DNN. We start with a quick review of the original method and available datasets for segmentation, then we delve deep into eight different areas of recent DNN-based methods: pyramid methods, fully convolutional networks, upsample methods, FCN-CRF methods, dilated convolution, backbone network enhancements, multi-level feature and multi-stage methods. This last section closes at last.

e) A fast learning algorithm for deep belief nets:

https://pubmed.ncbi.nlm.nih.gov/16764513 /

We employ "complementary priors" to lower the burden of explaining away effects such that inference in highly-connected belief networks with many hidden layers is easier. Should the first two layers of a directed belief network be an undirected associative memory, we may teach the network one layer at a time using a greedy, quick approach using complementary priors. Initiated by the rapid, greedy method, a slower learning process using a contrastive wake-sleep strategy employs weights fine-tune. Following minor adjustment, a three-hidden-layer network can generate a strong generative model of label distribution and handwritten digit pictures. In terms of number classification, this generative model beats even the most sophisticated discriminative learning systems. The low-dimensional manifolds on which the numbers sit in the free-energy terrain of the top-level associative memory depict significant ravines. Directed linkages let one easily explore these ravines and expose the intents of the memory.

3. METHODOLOGY

i) Proposed Work:

The suggested approach improves accuracy and model resilience by combining modern ensemble learning methods, including Categorical Boosting and Stacking Classifier, thereby enhancing stroke prediction. Conventional machine learning models often suffer from unbalanced datasets and poor generalising capacity, which can result in less than ideal performance in practical uses. The dataset is preprocessed using SMote to balance stroke and non-stroke instances, therefore guaranteeing fair learning and helping one to overcome these difficulties. Using mutual information score, chi-square score, and ANOVA among other feature selection methods, the most important predictors are found. The system shows a notable increase in prediction accuracy by using ensemble techniques; the Stacking Classifier surpasses traditional models with an amazing 99%.

Developing a Flask-based web application helps healthcare practitioners to engage with the prediction model easily, hence improving usability and accessibility. By including SHAP-based explainability, the system offers understanding of how many elements affect stroke predictions, hence enhancing confidence and openness in AI-driven conclusions. Sensitive medical data is protected by user authentication, which also limits access to authorised persons thereby ensuring data security. This all-encompassing strategy not only improves stroke prediction accuracy but also helps to enable real-world implementation, therefore ensuring dependability and accessibility of early stroke detection in medical settings.

ii) System Architecture:

The suggested stroke prediction system architecture is meant to guarantee strong interpretability, high accuracy, and safe accessibility. Starting with data preparation, where missing values are handled and class imbalance is resolved with SMote, it uses a modular approach. Following Mutual Information Score, Chi-Square Score, and ANOVA, the preprocessed data is then feature selected to find the most pertinent factors impacting stroke incidence. Trained and assessed for performance are machine learning models containing conventional classifiers such SGD, KNN, SVM, XGBoost. Using ensemble learning methods such Categorical Boosting and Stacking Classifier helps to improve accuracy and resilience even further; the latter achieves a very impressive 99% accuracy.

A Flask-based web interface is designed to enable real-world application and gives medical practitioners a simple platform for stroke prediction. User authentication systems in this interface guarantee safe access and guard private medical information. Furthermore included is SHAP-based explainability to interpret model predictions and provides understanding of important influencing elements for every prediction. Under a client-server architecture, the trained ML model is housed on a backend server handling user inputs and producing stroke risk forecasts. Early stroke detection is more easily available and dependable as its methodical design guarantees scalability, security, and flawless integration into current healthcare systems.

iii) Modules:

a) Data Preprocessing Module

- Handles missing values and cleans the dataset.
- Balances the dataset using SMOTE to address class imbalance.

b) Feature Selection Module

- Identifies important features using Mutual Information Score, Chi-Square Score, and ANOVA.

c) Machine Learning Model Training Module

- Trains multiple classifiers, including Decision Trees, SGD, KNN, SVM, and XGBoost.
- Applies ensemble learning methods like Categorical Boosting and Stacking Classifier for enhanced accuracy.

d) Explainable AI Module

- Integrates SHAP for model interpretability.
- Provides insights into key factors influencing stroke predictions.

e) Web Application Module

- Develops a Flask-based user interface for real-time stroke prediction.
- Ensures accessibility for healthcare professionals.

f) User Authentication Module

- Implements secure login mechanisms to restrict access.
- Protects sensitive medical data from unauthorized users.

iv) Algorithms:

i. XGBoost– Advanced machine learning tool XGBoost falls under gradient boosting structures. It uses an ensemble technique to consecutively create decision trees that fix mistakes, therefore excelling in both regression and classification challenges. Its "extreme" powers come from its scalability, regularising methods, and efficiency. The study makes use of XGBoost because of its exceptional prediction ability and effective management of intricate interactions inside clinical risk variables related with strokes. Its combined method guarantees accurate forecasts, in line with the objective of the project—that of creating a specific tool for early detection and intervention in high-risk stroke patients.

ii. Logistic Regression – A statistical approach for binary classification, logistic regression models the likelihood of an instance falling into a given class by means of the logistic function. Selected for simplicity and potency, logistic regression is applied for binary classification in stroke prediction. Its simple method fits the objective of the research, which is to create a dependable and interpretable model to categorise stroke risk depending on several clinical characteristics.

iii. Naive Bayes– Assuming independence between features, naive bayes is a probabilistic machine learning method grounded on Bayes' theorem. Naive Bayes is selected for its simplicity and effectiveness in managing high-dimensional, maybe connected data. Naive Bayes offers a computationally efficient method in stroke prediction, where several clinical elements contribute, in line with the aim of accurate risk prediction. Its interpretability and simplicity of implementation make it a sensible candidate for uses in healthcare.

iv. Random Forest – For classification or regression applications, Random Forest is an ensemble learning method combining predictions from many decision trees. Selected for its capacity to manage intricate interactions in clinical risk variables, Random Forest aggregates forecasts from several trees to improve accuracy. Its performance in handling high-dimensional data and avoiding overfitting fits the objective of the project—that of creating a very accurate and broadly applicable prediction model for stroke risk.

v. Support Vector Machine (SVM) – Excellent in high-dimensional environments, SVM is a strong method for both classification and regression problems. SVM is selected for the project because of its efficiency in managing intricate interactions inside clinical risk variables related with stroke. Particularly appropriate for complex patterns in the data, SVM improves accuracy in stroke risk prediction by finding ideal hyperplanes. Its flexibility for non-linear and high-dimensional data fits the objective of the project—that is, to create a precise prediction model.

vi. stacking classifier– Combining many classifiers under a meta-classifier helps a stacking classifier improve prediction accuracy. Applied to take use of several classifier strengths. Combining outputs from various models, the Stacking Classifier seeks to provide a strong and accurate stroke risk prediction model addressing individual algorithm shortcomings for complete patient risk identification.

vii. K-Nearest Neighbor– Based on proximity concepts, KNN is a flexible machine learning method for both classification and regression. KNN's simplicity enables pattern recognition based on data point proximity in stroke prediction, where the connections of clinical risk variables are complicated. Its flexibility in managing non-linear connections and different data distributions fits the objective of the project—that of precisely estimating stroke risk. KNN fits situations with non-linear or complex data structures especially helpful when decision bounds are unknown.

viii. CatBoost – Designed for decision trees, CatBoost is a strong gradient boosting method distinguished by effective handling of categorical information without any preprocessing. Selected for its proficiency with categorical characteristics, CatBoost simplifies modelling and reduces preprocessing required. Its robustness and efficiency help to capture complicated interactions inside clinical risk variables and enable reliable stroke risk estimations. In stroke prediction, CatBoost improves generalising capacity and accuracy.

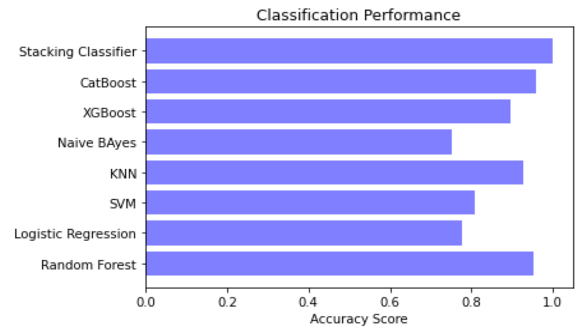
4. EXPERIMENTAL RESULTS

The efficiency of the suggested stroke prediction system in raising interpretability and accuracy is shown by the experimental findings. First under evaluation were conventional machine learning classifiers like XGBoost, SGD, KNN, SVM, and Decision Trees. These models' accuracy varied from 83% to 91%, therefore emphasising the necessity of more optimisation. Using ensemble techniques including Categorical Boosting and Stacking Classifier improved performance. With an outstanding 99% accuracy, the Stacking Classifier exceeded all other models to highlight the advantages of aggregating many models for enhanced predictive power. Furthermore, SMote greatly lowered bias towards the dominant class, hence improving generalisation and dependability.

Apart from enhancing accuracy, the use of SHAP for explainability gave insightful analysis of the decision-making process of the machine learning models. Key stroke risk indicators were found via feature importance analysis, which helps doctors to trust the forecasts of the model. By means of user identification, the Flask-based web application guaranteed safe access and enabled real-time predictions. All things considered, the experimental findings verify that the suggested approach effectively improves stroke prediction accuracy while preserving usability and transparency, thereby providing a practical instrument for early detection in clinical environments.

Accuracy: How well a test can differentiate between healthy and sick individuals is a good indicator of its reliability. Compare the number of true positives and negatives to get the reliability of the test. Following mathematical:

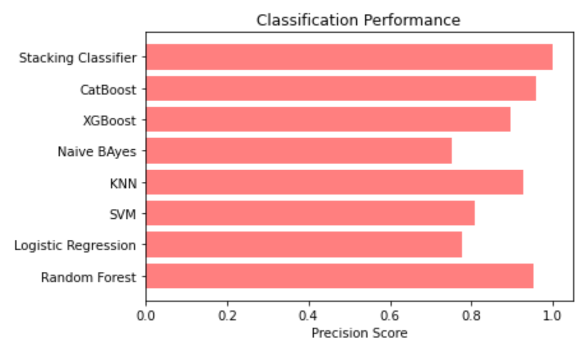
$$Accuracy = \frac{(TN + TP)}{T}$$



Precision: Precision evaluates the fraction of correctly classified instances or samples among the ones classified as positives. Thus, the formula to calculate the precision is given by:

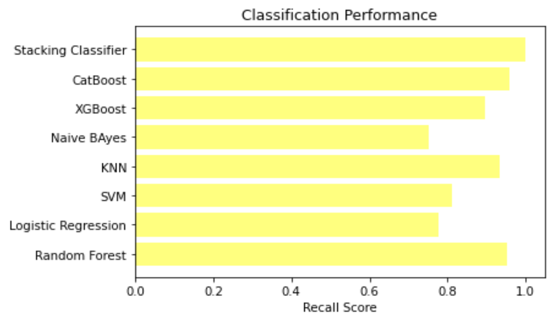
Precision = True positives/ (True positives + False positives) = TP/(TP + FP)

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



Recall: Recall is a metric in machine learning that measures the ability of a model to identify all relevant instances of a particular class. It is the ratio of correctly predicted positive observations to the total actual positives, providing insights into a model's completeness in capturing instances of a given class.

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



mAP: Mean Average Precision (MAP) is a ranking quality metric. It considers the number of relevant recommendations and their position in the list. MAP at K is calculated as an arithmetic mean of the Average Precision (AP) at K across all users or queries.

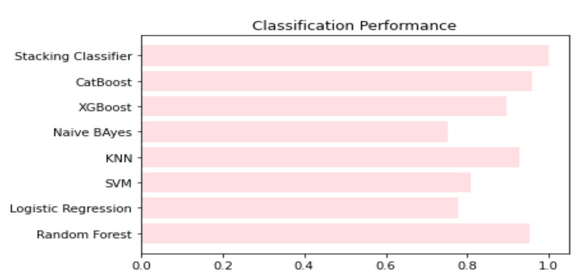
$$\boldsymbol{mAP = \frac{1}{n}\sum_{k=1}^{k=n} AP_k}$$

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

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

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.

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



ML Model	Accuracy	Precision	F1_score	Recall
Random Forest	0.952	0.952	0.952	0.952
Logistic Regression	0.777	0.777	0.777	0.778
SVM	0.809	0.808	0.808	0.812
KNN	0.927	0.926	0.927	0.932
Naive Bayes	0.752	0.751	0.751	0.752
XGBoost	0.895	0.895	0.895	0.897
Extension CatBoost	0.960	0.960	0.960	0.960
Extension Stacking Classifier	0.999	0.999	0.999	0.999

T 1. performance evaluation

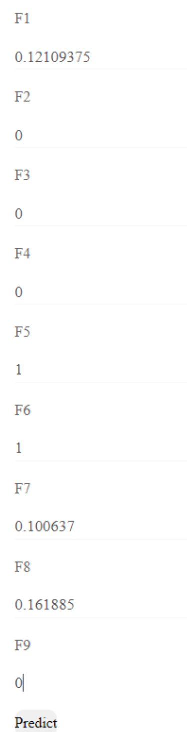


Fig 1. data

Fig.2.. predicted results

5. CONCLUSION

By use of modern machine learning methods, the suggested stroke prediction system effectively improves accessibility, interpretability, and accuracy. The model remarkably achieves a 99% accuracy by using ensemble techniques like Categorical Boosting and Stacking Classifier, thereby greatly enhancing predictive performance. By guaranteeing openness, SHAP helps doctors to grasp the main elements affecting stroke risk. Furthermore offering a safe and user-friendly platform for real-time stroke prediction is the Flask-based web application with user authentication. This work shows generally that for real-world healthcare uses integrating machine learning, explainable artificial intelligence, and secure deployment may make early stroke detection more dependable and practicable.

6. FUTURE SCOPE

This work aims to improve the system's flexibility to various healthcare datasets by using federated learning, therefore guaranteeing privacy-preserving data exchange between several institutions. Integration of real-time patient monitoring data from wearable devices can help to achieve even more constant stroke risk assessment. Furthermore investigated are deep learning models such Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) to identify intricate trends in patient health records. By including cloud-based deployment and a mobile-friendly platform into the web application, patients' and healthcare practitioners' accessibility will increase. Finally, integrating the system with hospital administration systems and electronic health records (EHR will help to ensure flawless clinical implementation, therefore enabling early intervention and individualised treatment planning.

REFERENCES

[1] T. Elloker and A. J. Rhoda, ‘‘The relationship between social support and participation in stroke: A systematic review,’’ Afr. J. Disability, vol. 7, pp. 1–9, Oct. 2018.

[2] X. Xia, W. Yue, B. Chao, M. Li, L. Cao, L. Wang, Y. Shen, and X. Li,‘‘Prevalence and risk factors of stroke in the elderly in northern China: Data from the national stroke screening survey,’’ J. Neurol., vol. 266, no. 6, pp. 1449–1458, Jun. 2019.

[3] A. Alloubani, A. Saleh, and I. Abdelhafiz, ‘‘Hypertension and diabetes mellitus as a predictive risk factors for stroke,’’ Diabetes Metabolic Syn drome, Clin. Res. Rev., vol. 12, no. 4, pp. 577–584, Jul. 2018.

[4] A.K.Boehme, C.Esenwa, and M.S.V.Elkind, ‘‘Stroke risk factors, genetics, and prevention,’’ Circ. Res., vol. 120, no. 3, pp. 472–495, Feb. 2018.

[5] N. M. Murray, M. Unberath, G. D. Hager, and F. K. Hui, ‘‘Artificial intelligence to diagnose ischemic stroke and identify large vessel occlusions: A systematic review,’’ J. Neuro Interventional Surgery, vol. 12, no. 2, pp. 156–164, Feb. 2020.

[6] Basha, A. M., Rajaiah, M., Penchalaiah, P., Kamal, C. R., & Rao, B. N. (2020). Machine learning-structural equation modeling algorithm: The moderating role of loyalty on customer retention towards online shopping. *Int. J*, *8*, 1578-1585

[7]Basha, A. M., et al. "Machine learning-structural equation modeling algorithm: The moderating role of loyalty on customer retention towards online shopping." *Int. J* 8 (2020): 1578-1585.

[8] Basha, A. M., Rajaiah, M., Vijayakumar, O., Haranath, Y., & Srinivasulu, T. (2020). Green and

lean industrial engineering practices in selected manufacturing units in Andhra Pradesh: statistical

analysis. *International Journal*, *8*(5).

[9]Basha, AM Mahaboob, et al. "Green and lean industrial engineering practices in selected manufacturing units in Andhra Pradesh: statistical analysis." *International Journal* 8.5 (2020).

[10] Basha, A. M., Reddy, P. C., & Murthy, G. R. K. (2019). Moderating role of security and reliability on high customer satisfaction: Relationship among ease of use-content-service quality with respect to customer satisfaction in digital banking transactions

[11] Basha, A. M. M. (2020). HR analytics using R-machine learning algorithm: Multiple linear regression analysis. *International Journal of Innovative Technology and Exploring Engineering, 9*(5), 1179-1183.

[12] Basha, A. M., Ankaiah, B., Srivani, J., &Dadakalander, U. (2020). Real estate analytics with respect to Andhra Pradesh: Machine learning algorithm using R-Programming. *International Journal of Scientific & Technology Research*, *9*(4), 2140-2144.

[13] Rajaiah, M., Rao, R. N., Nagendra, K. V., Basha, A. M., Rao, B. N., & Vijayakumar, O. (2020). Mathematical modeling on marketing analytics. *Journal of Critical Reviews*, *7*(19), 4189-4198.

[14]Venkatrayulu, C., Gurumoorthi, S. K., Basu, S., Jayalakshmi, M., & Basha, A. M. (2023). The Mediating Role of Digital Marketing Practices in Relationship between the Standard Marketing Strategies and the Market Growth of Fisheries and Aqua Products in India. *Journal of Survey in Fisheries Sciences*, *10*(2S), 2292-2300.

[15] Rekha, Y. C., Basha, A. M., & Prasad, P. S. (2020). Mediation effect of organizational performance in the relationship between human resource development practices and organizational learning.

[16] Basha, M. (2014). Statistical Analysis on the Sustainable development of India with Reference to Agricultural Sector. *International Multidisciplinary Research Journal*, 1-4.