Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 3, No. 5, March : 2024

COMPUTATIONAL TECHNIQUES FOR CANCER DETECTION AND RISK EVALUATION

Amiya Kumar Sahoo, Aparna Rajesh A, Prakash Kumar Dehury, Er. Anaryami Badu, Dr. Bimal Sarangi, Priya Chandan Satapathy, Department of Computer Science & Engineering, Aryan Institute of Engineering and Technology BPUT, Odisha, India.(dramiya79@gmail.com)

INTRODUCTION

Cancer presents an enduring and formidable global public health challenge, demanding a continuous search for innovative strategies focused on early detection and precise risk assessment. Its pervasive nature and severe consequences underscore the critical need for fresh and advanced approaches. In response, our research endeavours to introduce a pioneering solution that merges the capabilities of data science and Internet of Things (IoT) technology to transform lung cancer prediction and evaluation. Despite notable advancements in medical science and unwavering efforts in public health, lung cancer remains a significant contributor to cancer-related deaths on a global scale. Its insidious nature, often remaining asymptomatic until reaching advanced, and frequently untreatable stages, emphasizes the immediate urgency of early detection. In this context, our study establishes a novel paradigm, with a primary focus on utilizing a meticulously curated dataset encompassing 1000 individuals. This dataset serves as the foundation of our research, including a diverse range of patient attributes such as demographics, lifestyle factors, medical history, and data collected from IoT devices. These multifaceted data sources provide the basis for constructing our predictive models, equipping them with the necessary information for making more informed and precise assessments.

At the heart of our innovative approach lie cutting-edge machine learning models, including Support Vector Machines (SVM), Naïve Bayes Multinomial (NBM), Meta Bagging, PART, and Random Forest (RF). The integration of these advanced models substantially enhances the accuracy of our predictions, providing healthcare professionals and individuals with a potent tool for making critical decisions pertaining to lung cancer detection and risk assessment. A distinctive feature of our research is the seamless integration of IoT technology for real- time and continuous health monitoring. This dynamic approach transcends the limitations of intermittent snapshots, ensuring an ongoing and adaptive evaluation of lung cancer risk. By synergizing the strengths of robust data analysis with the capabilities of the Internet of Things, our research aims not only to identify lung cancer in its earliest, most treatable stages but also to offer personalized risk assessments tailored to each individual's unique health profile. Our study reveals the exceptional performance of the PART model, achieving an impressive accuracy rate of 95%. This surpasses the performance of other models, including NBM (67%), RF (93%), SVM (92%), and Bagging (86%). These results underscore the effectiveness of our approach and its potential to revolutionize the landscape of lung cancer detection and risk assessment. In summary, our research offers a groundbreaking opportunity to advance the field of early lung cancer detection, empowering individuals to take proactive measures for their health and contributing to improved patient

Keywords: cancer prediction, machine learning, support vector machines, Naive Bayes

1. LITERATURE REVIEW

Here introduced an innovative approach to data classification by leveraging neural networks and exploring various soft computing models. Their study showcases the strategic application of neural networks in a novel manner, providing insights into effective data classification methods. Swain. S. et al. [2] illness prognosis, focusing on diabetes and heart diseases in the elderly. It emphasizes the potential of IoT and machine learning to aid in diagnosis and prevention. Additionally, it highlights the disconnect between healthcare and technology and suggests that new technologies, including soft computing models and Artificial Neural Networks, can significantly benefit heart disease treatment. same team focused on smart weather prediction for Delhi. They employed a variety of machine learning

techniques, including Random Forest, Decision Tree, Support Vector Machine, Neural Network, Adaboost, Xgbost, Gradient Boosting, Naïve Bayes, and Logistic Regression. Their findings indicated that Random Forest outperformed the other machine learning models, enhancing the accuracy of weather prediction. Furthermore, Jayasingh et al. [4] exploration of hybrid soft computing models for weather prediction revealed that these hybrid approaches surpassed traditional soft computing models in accuracy and error parameter metrics.
Here delved into the domain of SMS fraud detection using various machine learning methods, such as Random Forest, Naïve Bayes, KNN, Decision Tree, Ada Boost, and Support Vector Machine. Their study showcased the superiority of Random Forest in achieving the highest accuracy among the methods considered, contributing to more effective fraud detection.

2. RESEARCH DESIGN

2.1. Architecture of ML

The architecture of a Machine Learning (ML) system comprises a series of pivotal components and stages. It commences with data acquisition and storage, where diverse data sources are aggregated and stored for easy retrieval. Subsequently, data preprocessing ensues, encompassing data cleansing, transformation, and feature engineering. The extraction and selection of pertinent features become imperative to hone in on essential information. The selection and training of models constitute critical phases, wherein various algorithms like decision trees and neural networks are employed. Model evaluation serves to ensure performance aligns with expectations. Hyperparameter tuning fine-tunes model effectiveness. Transitioning to deployment within production environments occurs next, underpinned by continuous monitoring, maintenance, and scalability considerations. Upholding privacy and security protocols is integral, while enhancing interpretability and user interfaces facilitates user interaction. Ongoing feedback loops iteratively enhance model performance. Within this meticulously designed architecture, ML systems can be customized for specific applications and domains, all while upholding efficiency, security, and ethical principles as shown below figure - 1.

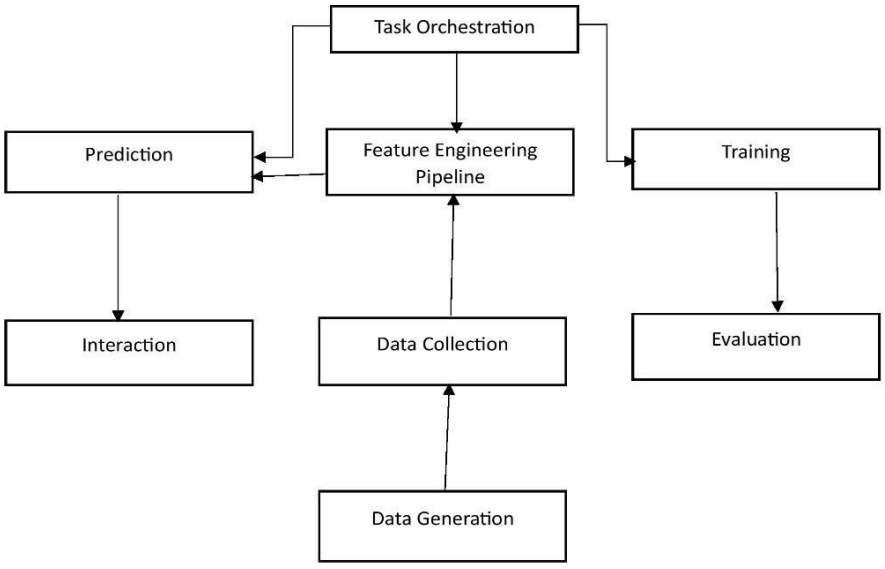

Figure – 1 : Work flow diagram of Machine Learning

2.1.1. How ML differ from traditional Programming

Traditional programming relies on explicit rules written by developers, whereas ML is driven by data, learning patterns and making predictions. ML is particularly suited for complex, data-rich tasks like image recognition. In contrast, traditional programming excels in well-defined roles, such as database management, where explicit rules suffice. Moreover, in ML, the model training phase includes assigning new datasets to train the algorithm for predictions, showcasing the adaptability of ML systems. The choice between the two approaches depends on problem complexity and data availability

as shown kin Figure – 2 and 3.

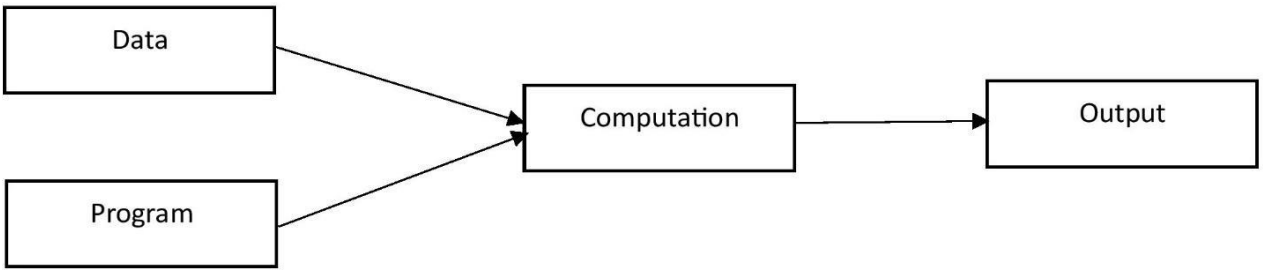

Figure – 2 : Traditional Program work flow

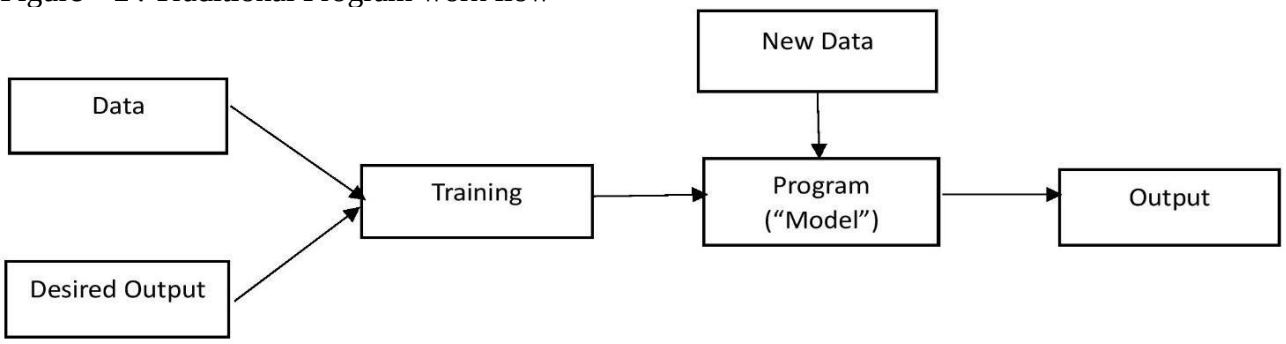

Figure – 3 : Work flow of Machinelearning Processing

2.2. Voting Model

A voting model in the realm of machine learning is a type of ensemble model that amalgamates the forecasts of multiple individual models to formulate a conclusive prediction. This amalgamation can be achieved through various means, such as computing a weighted average of the predicted probabilities from these individual models or selecting the class that is predicted by the majority of these models. The primary objective of using voting models is to enhance the accuracy and robustness of machine learning systems. By amalgamating the predictions from multiple individual models, voting models play a pivotal role in mitigating the risks associated with overfitting and enhancing the system's resilience against noise and outliers present in the data. In the context of our research, we have incorporated five distinct machine learning models: Support Vector Machines (SVM), Naïve Bayes Multinomial (NBM), Meta Bagging, PART, and Random Forest (RF), with the aim of predicting lung cancer.

2.3. Navie Bayes Multinomial

A popular supervised machine learning technique for text categorization applications is Naive Bayes Multinomial (NBM). This kind of classifier belongs to the Naive Bayes family, which is a group of algorithms founded on the Bayes theorem. The Bayes theorem is a mathematical formula that lets us figure out how likely it is for an event to happen based on the knowledge of an earlier event.
Formula can be used to calculate:
P(B|A) / P(B) * P(A) = P(A|B)
where prior probability of B = P(B) Prior Probability of Class A (P(A))
P(B|A) is the probability of predictor B given class A.

2.4. Support Vector Machine

For problems involving regression and classification, Support Vector Machine (SVM) is a reliable supervised machine learning technique. In order to maximize the margin between data points, it finds a hyperplane that best divides them into different classes. SVM is proficient in high-dimensional spaces, has the capability to manage non-linear data using kernel functions, and exhibits resilience against overfitting. It finds extensive application in various fields, including image classification, text categorization, and bioinformatics. SVM's proficiency in dealing with intricate datasets and determining optimal decision boundaries positions as shown in Figure - 4.
SVM divided into two types, Based on the training sets,

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 3, No. 5, March : 2024

- Linear SVM – A linear line makes it simple to segregate data points.
- Non-Linear SVM – It is difficult to divide data points using a straight line.

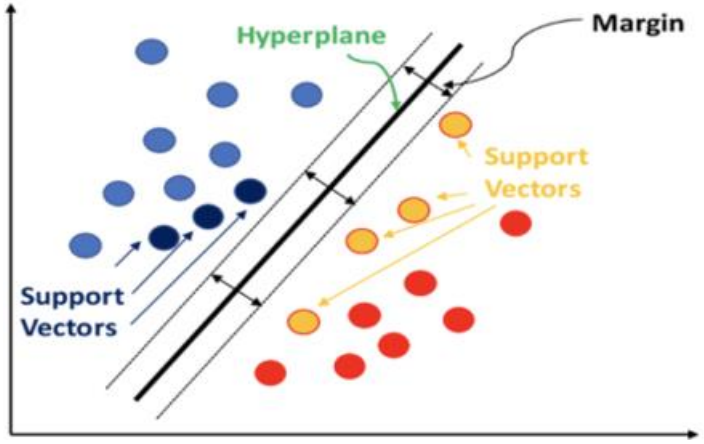

Figure – 4: Working principle of SVM

2.5. Partial Decision Tree (PART)

PART is also tailored for classification tasks. It employs a decision tree approach to construct concise and intelligible trees by emphasizing the selection and refinement of attributes and values that offer maximum classification insight. Beginning with a complete decision tree, PART trims it by eliminating branches with limited impact on classification precision. It yields a more transparent and interpretable decision tree, a valuable asset when comprehensibility is paramount. PART finds application across domains, including data mining and medical diagnosis, where lucid decision logic is essential.

2.6. Random Forest (RF)

Random Forest (RF) is a potent ensemble learning method extensively utilized in machine learning. It functions by building multiple decision trees during training and amalgamating their predictions for enhanced accuracy and robustness. It exhibits proficiency in handling both classification and regression tasks, managing high-dimensional data, and mitigating overfitting. Its adaptability to diverse data types, coupled with the randomness in tree construction and feature selection, grants RF resilience against noise and outliers. RF's versatility finds application across a spectrum of domains, due to its exceptional predictive capabilities and flexibility we can use it in encompassing image classification, bioinformatics, medical and financial analysis as shown in Figure - 5.

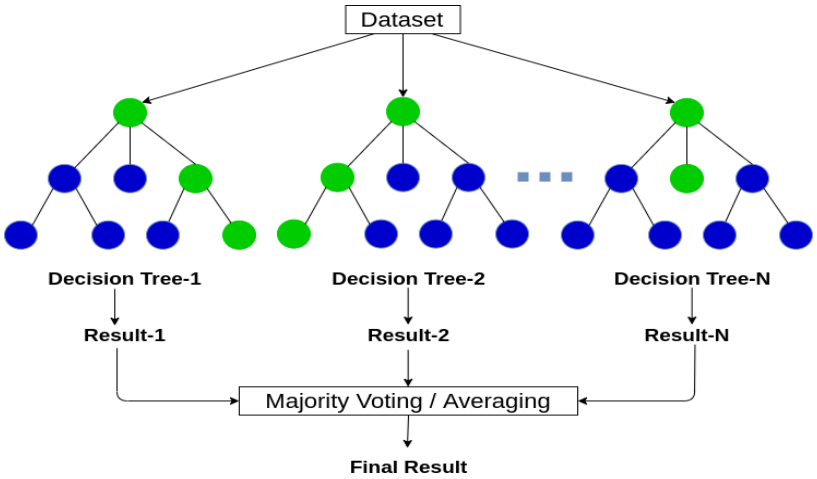

Figure – 5 :Graphical representation of RF

2.7. Bagging

Bagging, which stands for Bootstrap Aggregating, is an ensemble learning method that helps decision trees and other machine learning models become more accurate and stable. It operates by bootstrapping (random sampling with replacement) to create several subsets of the training data, then training a base model on each subset.. The final prediction is often an average or a majority vote of the predictions

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 3, No. 5, March : 2024

made by the individual models. Bagging reduces overfitting, enhances model robustness, and is particularly effective when applied to unstable or high- variance algorithms as shown in Figure – 6.

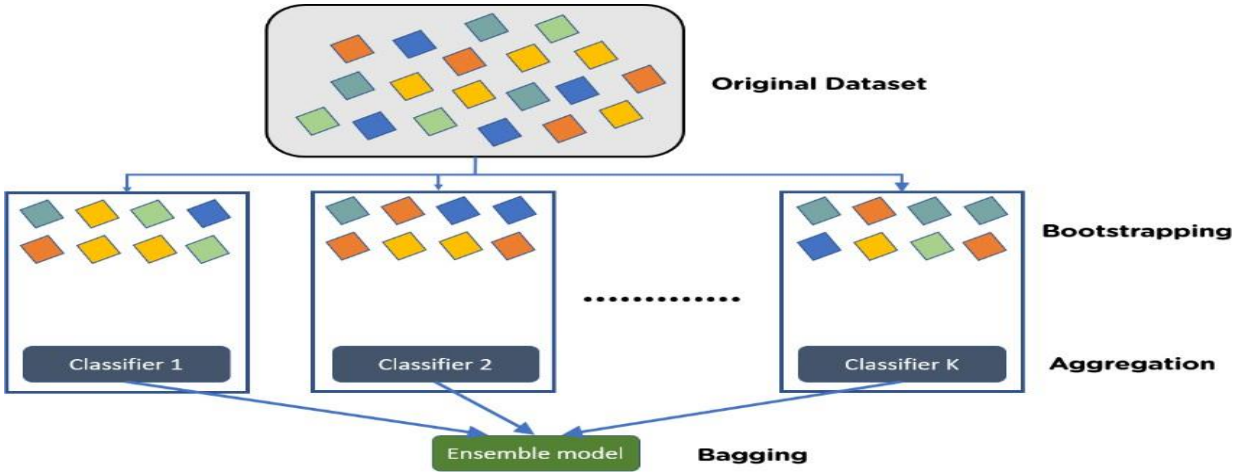

Figure – 6: Working principle of Bagging Model

3. RESEARCH METHODOLOGY

3.1. Flow of work

The flow chart of our proposed work is shown below Figure – 7

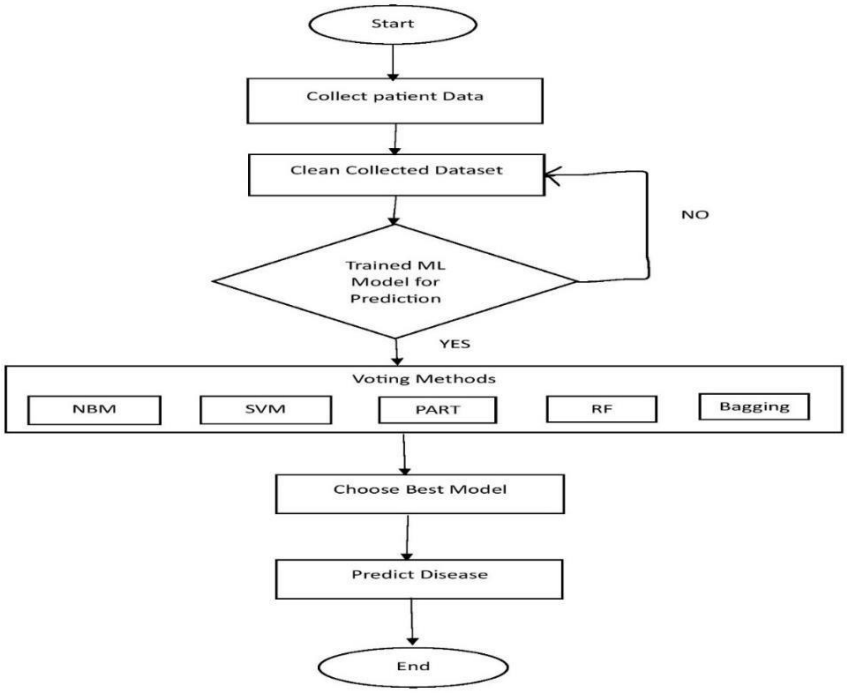

Figure – 7 : Flow Chart of propoed work

3.2. Data collection

we gathered a dataset comprising 1000 individuals diagnosed with lung cancer, sourced from a Kaggle dataset. This dataset includes various parameters such as age, gender, lifestyle factors, and medical history, which encompass attributes like air pollution, smoking, chronic lung disease, and more. Our primary goal is to utilize this data to predict the stages of lung cancer. By employing machine learning techniques and data analysis, we aim to develop a predictive model that can assist in the early diagnosis and staging of lung cancer, ultimately contributing to more effective treatment and patient care.

3.2.1. Clean dataset

By using the mean, mode, and median we clean the collected dataset, which helps ensure the dataset

is more complete and ready for analysis by minimizing the impact of missing or outlier values.

- Mean is the average of all the values in a data set. It is calculated by adding up all the values and dividing by the number of values.

For example,
if the data set is {1, 2, 3, 4, 5},
the mean is (1 + 2 + 3 + 4 + 5) / 5 = 3

- Mode is the most frequent value in a data set. For example,

in the data set {1, 2, 3, 4, 5}
the mode is 3 because it appears the most times.

- Median is the middle value in a data set when the values are arranged in order from smallest to largest. For example,

in the data set {1, 2, 3, 4, 5}
the median is 3 because it is the middle value when the data is arranged in order.

4. COMPARISON ANALYSIS

We comparing used machine learning models in our research i.e., Support Vector Machines (SVM), Naïve Bayes Multinomial (NBM), Meta Bagging, PART, and Random Forest (RF) based on various evaluation metrics is a common practice in data analysis and model selection. Here are some key evaluation metrics used for model comparison as shown in Figure -8.

- Predictive Accuracy
- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)
- Relative Absolute Error (RAE)
- Relative Root Mean Squared Error (RRSE)

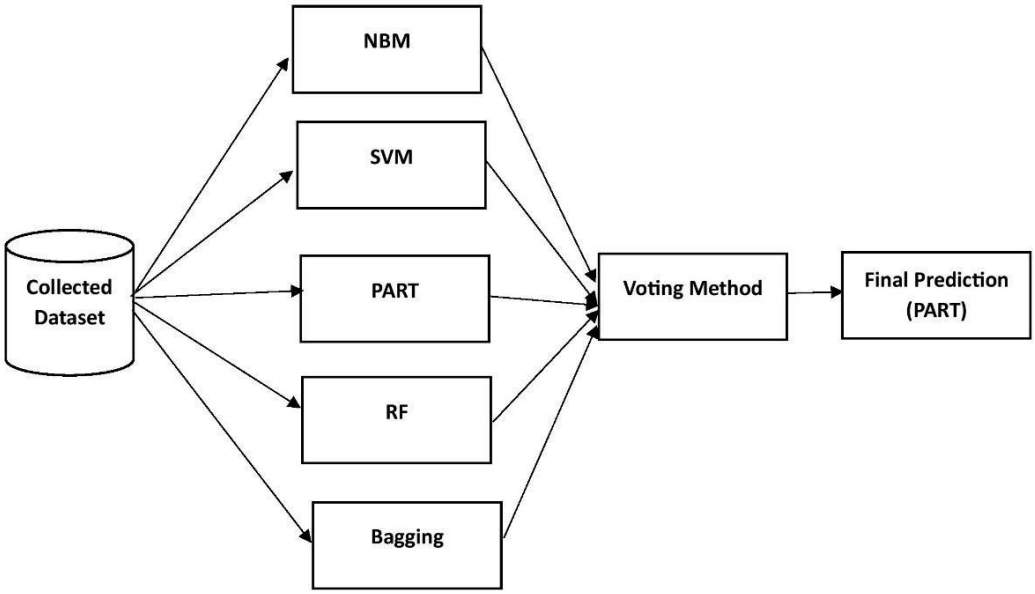

Figure – 8 :Work flow of Voting Model

It achieved individual accuracy scores for each model. Here are the accuracy scores for each model:
Navie Bayes Multinomial (NBM): 67.3%
Support Vector Machine (SVM): 92.96% PART: 95.23%
Random Forest (RF): 93.17% Meta. Bagging: 86.23%
Here we found PART gives best accuracy than other four models.
Here we split our collected dataset into five equal set and apply models individually and store the error value in tables 1,2,3,4 and 5. The pictorial representations of comparisons of different models are shown in Figure 9,10,11,12 and 13.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 3, No. 5, March : 2024

Table - 1 Accuracy table of Voting model of NBM, SVM, PART, RF and Bagging over 5 equal divided datasets

Dataset	NBM	SVM	PART	RF	Bagging
1-200	68.34	97.98	97.48	98.22	91.95
200-400	69.69	93.93	98.98	98.98	90.4
400-600	72.73	93.93	95.95	97.97	94.44
600-800	73.73	95.45	98.48	97.97	93.43
800-1000	74.39	97.58	98.55	99.03	93.71

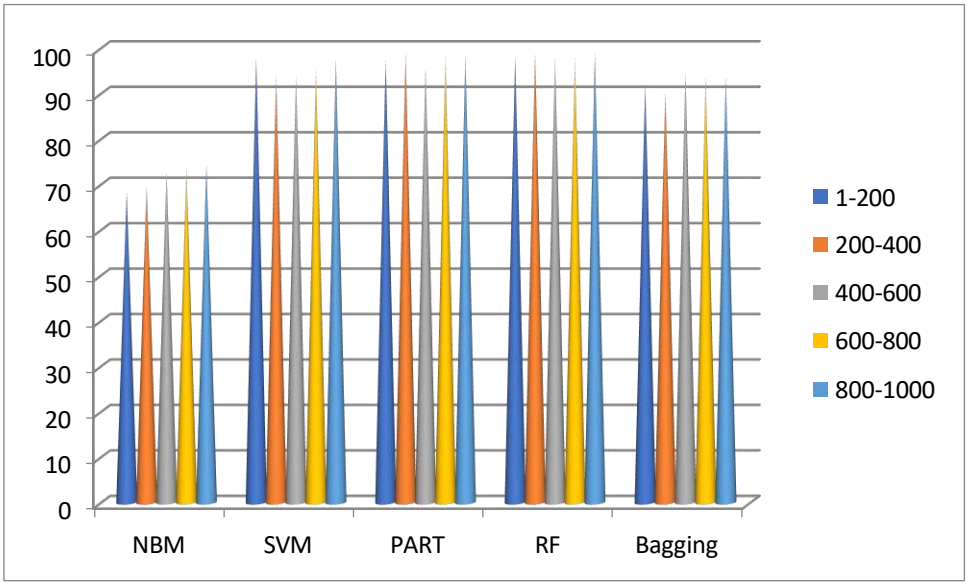

Figure – 9 Pictorial representation of comparison of NBM, SVM, PART, RF and Bagging on the basis of Accuracy

Table - 2 MAE table of Voting model of NBM, SVM, PART, RF and Bagging

Dataset	NBM	SVM	PART	RF	Bagging
1-200	0.21	0.13	0.01	0.03	0.09
200-400	0.19	0.04	0.007	0.03	0.08
400-600	0.19	0.04	0.02	0.04	0.07
600-800	0.17	0.03	0.01	0.03	0.07
800-1000	0.18	0.01	0.01	0.03	0.07

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 3, No. 5, March : 2024

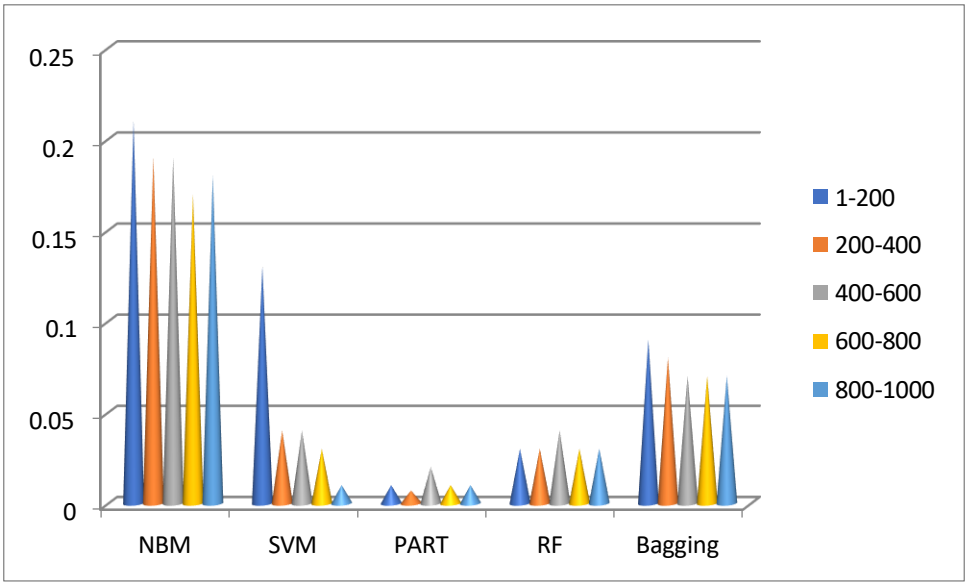

Figure – 10 Pictorial representation of comparison of NBM, SVM, PART, RF and Bagging on the basis of MAE

Table - 3 RMSE table of Voting model of NBM, SVM, PART, RF and Bagging

Dataset	NBM	SVM	PART	RF	Bagging
1-200	0.39	0.11	0.12	0.07	0.19
200-400	0.36	0.2	0.08	0.08	0.18
400-600	0.37	0.2	0.16	0.11	0.17
600-800	0.35	0.17	0.1	0.11	0.17
800-1000	0.35	0.12	0.09	0.08	0.16

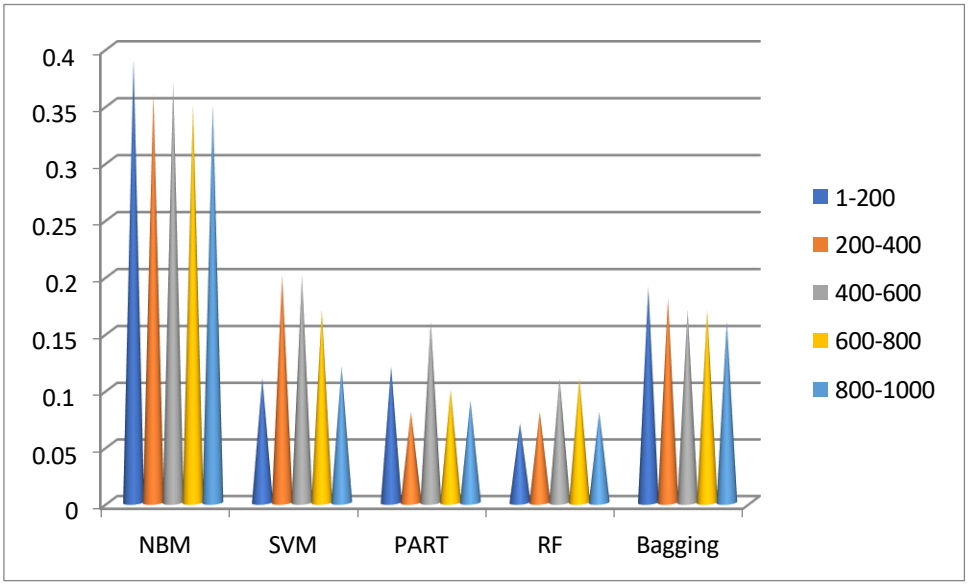

Figure – 11 Pictorial representation of comparison of NBM, SVM, PART, RF and Bagging on the basis of RMSE

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 3, No. 5, March : 2024

Table - 4 RAE table of Voting model of NBM, SVM, PART, RF and Bagging

Dataset	NBM	SVM	PART	RF	Bagging
1-200	49.19	3.02	3.93	7.24	21.22
200-400	45.65	9.28	1.69	7.44	19.74
00-600	44.16	9.29	6.36	9.39	18.24
600-800	40.22	6.86	3.18	8.03	16.74
800-1000	39.84	3.65	2.47	7.18	16.91

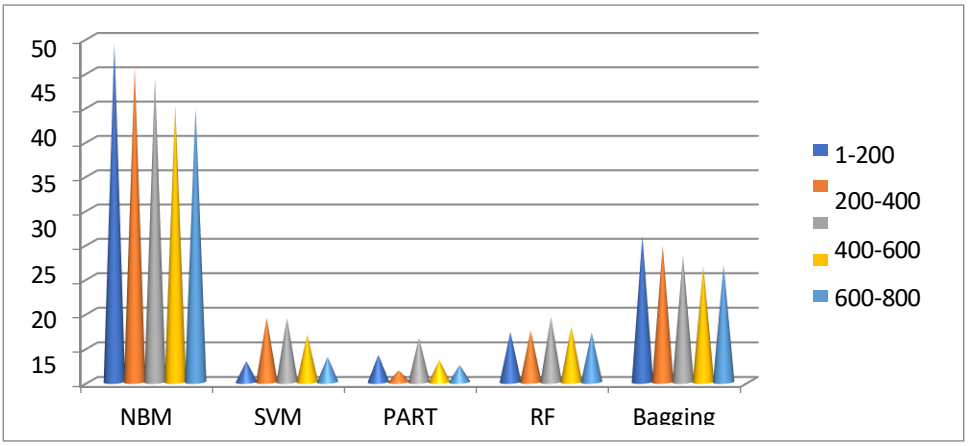

Figure – 12 Pictorial representation of comparison of NBM, SVM, PART, RF and Bagging on the basis of RAE

5. CONCLUSION

As we look to the future, our study paves the way for several exciting avenues of research. Firstly, the inclusion of more extensive and diverse datasets holds the promise of enhancing the robustness of our predictive models. Additionally, the integration of advanced deep learning techniques may further refine the accuracy and precision of lung cancer prediction. Exploring real-time IoT data streaming for continuous monitoring and immediate risk assessment opens the door to dynamic and proactive healthcare. Lastly, prioritizing the interpretability and explanation of our models is of utmost importance, ensuring that both the medical community and patients can place their trust in and reap the benefits of these predictive tools. Our journey in the fight against lung cancer is an ongoing one, and these forthcoming initiatives will undoubtedly propel progress in this vital field of research.

REFERENCES

1. Swain, S.., Behera, N.., Swain, A. K., Jayasingh, S. K.., Patra, K. J.., Pattanayak, B. K.., Mohanty, M. N.., Naik, K. D.., & Rout, S. S. (2023). Application of IoT Framework for Prediction of Heart Disease using Machine Learning. International Journal on Recent and Innovation Trends in Computing and Communication, 11(10s), 168–176. Retrieved from https://ijritcc.org/index.php/ijritcc/article/view/7616
2. Prusty, S.R., Sainath, B., Jayasingh, S.K., Mantri, J.K. (2022). SMS Fraud Detection Using Machine Learning. In: Udgata, S.K., Sethi, S., Gao, XZ. (eds) Intelligent Systems. Lecture Notes in Networks and Systems, vol 431. Springer, Singapore. https://doi.org/10.1007/978-981-19-0901-6_52.
3. Xu, Y., Hosny, A., Zeleznik, R., Parmar, C., Coroller, T., Franco, I., ... & Aerts, H. J. (2019). Deep learning predicts lung cancer treatment response from serial medical imaging. Clinical Cancer Research, 25(11), 3266-3275.
4. Avanzo, M., Stancanello, J., Pirrone, G., & Sartor, G. (2020). Radiomics and deep learning in lung cancer. Strahlentherapie und Onkologie, 196, 879-887.
5. Hyun, S. H., Ahn, M. S., Koh, Y. W., & Lee, S. J. (2019). A machine-learning approach using PET-based radiomics to predict the histological subtypes of lung cancer. *Clinical nuclear medicine*, *44*(12), 956- 960