JOURNAL OF NORTHEASTERN UNIVERSITY

Volume 26 Issue 01, 2023 ISSN: 1005-3026 https://dbdxxb.cn/ Original Research Paper

CLASSIFICATION OF RISKY CANCER PATIENTS USING ENHANCED MULT ALGORITHM

K. Divya Daniel[1], C. Shoba Bindu[2], P. Dileep Kumar Reddy[3]

[1]Department of CSE, JNTUA College of Engineering, Ananthapuramu, Andhra Pradesh, India divyadaniel12@gmail.com

[2]Department of CSE, JNTUA College of Engineering, Ananthapuramu, Andhra Pradesh, India shobabindhu.cse@jntua.ac.in

[3]Department of CSE, Narsimha Reddy Engineering College (Autonomous), Secunderabad, Telangana, India, dileepreddy503@gmail.com

ABSTRACT:

All cancers stem from underlying cellular DNA abnormalities and might appear at any time in a person's life. Because of the wide range of resulting genetic and phenotypic variations among afflicted individuals, the search for viable therapeutics is time-consuming and resource-intensive. Due to the rarity of samples and the abundance of input parameters, cancer datasets are notoriously difficult to work with when attempting to construct reliable predictors for classifying patients into risk groups. This article discusses four different types of cancer, lung cancer, breast cancer, pancreatic cancer and kidney cancer. To better understand how to categorize cancer patients into those at high and low risk, this research employs machine learning and deep learning algorithms for prediction, relying on a mix of supervised, unsupervised and self-supervised learning methodologies. This study's findings add to the growing body of evidence supporting the use of integrated learning algorithm and the combination of genetic and clinical data in the support of Clinical oncology decision-making.

INDEX TERMS: Cancer, Pancreatic cancer, renal cancer and machine learning or deep learning algorithms.

I. INTRODUCTION:

For fifty years, oncology research has profited enormously from the integration of engineering and the physical sciences. The convergence of these areas, accelerated by ML algorithms, may lead to innovative computer models for modelling complicated cancer systems, which might enhance treatment results while reducing costs. Because cancer is the second biggest cause of death globally accounting for nearly 10 million deaths in 2020 or nearly one in six deaths. Each year, approximately 400 000 children develop cancer. The number of Indians suffering from cancer is projected to increase to 29.8 million in 2025 from 26.7 million in 2021. All malignancies result from mutations in cellular DNA, yet both genotype phenotypes may vary. Correct identification of therapies is hampered as a result; this may increase the likelihood of more complicated, costly, and unsuccessful interventions. Overall survival has improved

Submitted: 11/12/2022 **Accepted**: 06/01/2023

because of advances in cancer treatment, but only a subset of individuals get any benefit. Due to this, it is important to find better ways to diagnose cancer and assess its risks, which motivates the search for novel indicators of treatment response. Therefore, the identification and collection of relevant markers is an essential initial step in creating unique, effective treatment strategies. Because to achieve rapid progress in biotechnology, the whole human genome can now be read in its entirety. Based on a small number of really insightful forecasts, this information directly affects whether it is feasible to develop individualized treatments.

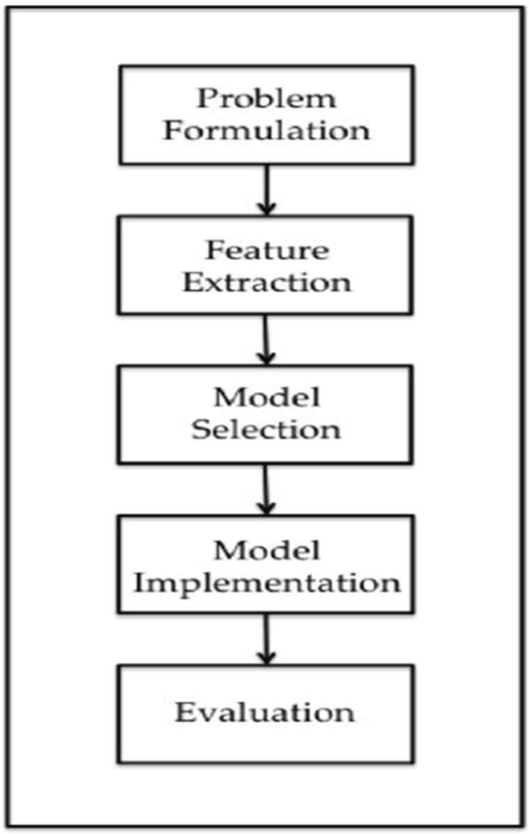

Figure 1: Overview of Architecture [1]

Huge quantities of genetic and clinical data on cancer have been gathered as a direct consequence of these developments. However, massive data sets provide a variety of statistical and computational challenges [1]. Some of the concerns include significant sparsity due to limited sample size and a large number of molecular characteristics and an unequal distribution of samples for each cancer occurrence.

The method has the following features: I it accepts various patient data types as input, allowing for consideration of various data representations; (ii) Using a feature selection step, it avoids noisy and pointless features; (iii) it finds latent information in the data by engineering features using unsupervised learning techniques; (iv) it employs an autoencoder to produce a more

robust representation of the data; and (v) it makes use of highly precise measurement techniques. Figure 1 depicts a high-level view of the plan's elements. Lung cancer, breast cancer, pancreatic cancer, and kidney cancer are the four forms of cancer that are covered in this essay. This research uses machine learning and deep learning algorithms for prediction, depending on a combination of supervised, unsupervised, and self- supervised learning approaches, to better understand how to categorise cancer patients into those at low and high risk. [9] The results of this study add to the growing body of research that demonstrates the effectiveness of integrated learning algorithms and the integration of genetic and clinical data in improving clinical oncology decision-making.

Due to the broad use of state-of-the-art forecasting models, these models are more effective when compared to the traditional methods.

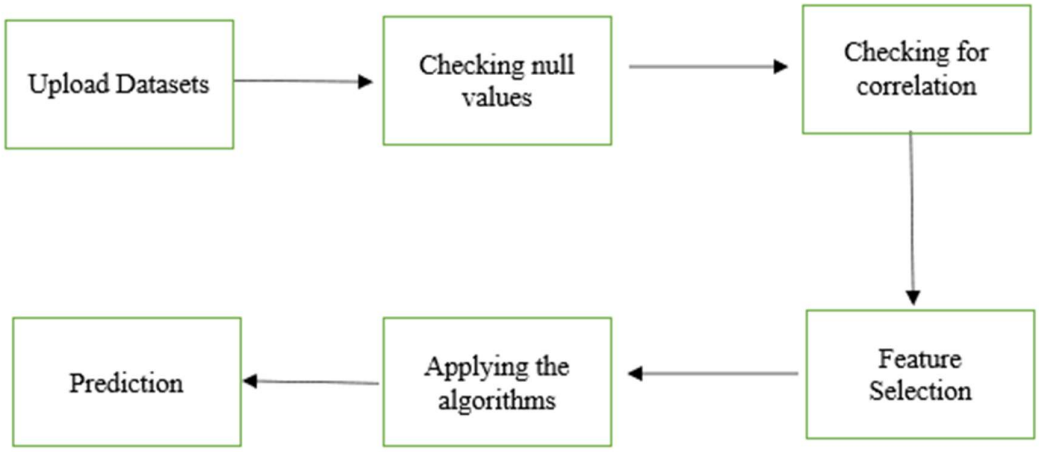

Figure .2: System Architecture

II. RELATED WORK

• Since the 1960s, using ideas from engineering and physics has benefited the discipline of cancer. Major advances in cancer diagnosis and treatment, as well as the quantitative knowledge of tumour genesis and progression, have all been attributed to engineers and physical scientists. Early studies on drug dispersion, cell cycle kinetics, and tumour development dynamics combined experimental and computer modelling. Thanks to developments in materials science, microfabrication, nanomedicine, microfluidics, imaging, and new NIH programmes like the National Institute of Biomedical Imaging and Bioengineering (NIBIB), Physical Sciences in Oncology (PSO), and the National Cancer Institute (NCI) Allia, the intersection of engineering, physics, and oncology has grown exponentially over the past ten years. [1] [2] The physical environment of the tumour, microfluidics, microfabrication, and cellular and molecular imaging are four areas where engineering and physical sciences have collided with cancer. We talk about the possibilities for engineering and physical sciences to be combined with oncology in the present and the future to develop new techniques for cancer research, diagnosis, and therapy. [3]

- The Wisconsin female breast cancer tumour data set is used as the study object, and the Random Forest and AdaBoost algorithms are integrated to create a breast cancer classification prediction model that can identify whether a tumour is benign or malignant. The Support Vector Machine, Logistic Regression, K-Nearest Neighbor, and Decision Tree algorithms are then tested against the model individually [4].

- The methods include decision trees, support vector machines, and artificial neural networks (ANNs) (DTs). While ML approaches can be used to understand how cancer develops, a high enough level of validity is needed to frequently use these techniques in clinical practice. This paper [5] reviews the ML & DL approaches used in cancer progression modelling. Most of the discussed predictions are connected to specific ML, input, and data sample supervision.

- Cancer therapy and prevention working together: Establishing a holistic structure while certain tumors have responded well to targeted therapies and following the growth of "personalized oncology," there are still significant obstacles to be addressed before this approach can be used on a large scale. Targeted therapies show promises; however, most patients relapse within a few months owing to the therapy's toxicity, high cost, or both. The existence of therapy resistant immortalized cells that adapted by employing alternative compensatory pathways is a have and major contributor to tumour relapses caused by genetic heterogeneity (i.e., pathways that are not reliant upon the same mechanisms as those which have been targeted). To get around these limitations, 180 scientists from across the globe banded up to study the potential of a "broad- spectrum" therapeutic technique that uses a low damaging dosage to target several pathways and processes all at once. Cancer hallmark phenotypes and the tumour microenvironment were used to evaluate each hallmark region, and a large number of high-priority targets (74 in total) were identified for adjustment to improve patient outcomes, taking into consideration the numerous components of relevant cancer biology.[6] More over two-thirds of the combinations suggested synergistic effects, whereas the remaining third were completely novel. Comparatively, mutually advantageous alliances accounted for 62.1% of the strategies, symbiotic partnerships for 13.9%, and hostile partnerships for 1.1%. These results suggest that, from a safety perspective, a broader approach is feasible. This novel approach has the potential to reduce recurrence rates, treat malignancies at later stages and of different kinds than those targeted by normal therapy, and come at a relatively modest cost. [2]

- Mutational fingerprints offer a rare chance to group tumour types with comparable evolutionary trajectories and prognoses. This signal was discovered using non-negative matrix factorization (NMF) methods applied to high throughput sequencing data from cancer genomes. The NMF paradigm is highly dimensional and nonconvex, making it difficult to use current state-of-the-art solutions based on optimization techniques. [7] There is a pressing need to know how many signatures are sufficient for a reliable representation of the data. Further research is needed to create effective algorithms for extracting mutational signatures from high-throughput data.

- A novel approach for the statistical estimate of mutational signatures is presented in this study using an empirical Bayesian treatment of the NMF model. Since our method treats the quantity of signatures as a model selection issue, little user input is required. Additionally, we offer two fresh ideas for assessing the mutational profile, both of which are highly therapeutically relevant. We examine both real-world and simulated data to demonstrate the efficacy of our strategy. The latter is utilised to compare our strategy to two others that are often cited in the literature and have the same NMF parametrization. Comparatively speaking, our method is more accurate and resilient to changes in the initial conditions. Even when other methods are unsuccessful, it can estimate the correct number of signatures. Results from actual data are consistent with what is known right now.

- The significance of genetic variation in carcinogenesis and progression Recent studies have shown extensive genetic heterogeneity across and even within individual tumours. This variety impacts key cancer pathways, which in turn generates phenotypic variation, further complicating attempts to create individualised therapies for cancer. The genetic variety of cancers is mostly attributable to chromosomal instability.[8] This instability increases the rate of mutation, which may have several effects on the evolution of cancer genomes. By investigating these mechanisms, we may get insight into normal patterns of tumour. All cancers stem from underlying cellular DNA abnormalities and might appear at any time in a person's life. Because of the wide range of resulting genetic and phenotypic variations among afflicted individuals, the search for viable therapeutics is time consuming and resource-intensive. Due to the rarity of samples and the abundance of input parameters, cancer datasets are notoriously difficult to work with when attempting to construct reliable predictors for classifying patients

into risk groups.

Additionally, due to the limited sample size and numerous input variables, cancer datasets are frequently sparse. Creating a challenging environment for the creation of accurate forecasts for risk categorization of patients has become a challenge.

METHODOLOGY

As shown in Figure 3. This paper is categorized into the following steps, where initially the data is gathered from public resources. This data is pre-processed to remove the null values so that effective usage is made out of the data. The next step includes labelling of the data as semi supervised techniques are implemented. After the data is processed and labelled, we work on the MuLT instances where multiple algorithms like random forest, logistic regression, SVM etc., and models like ResNet and DenseNet are executed. When the accuracy, precision, recall and F1 score reach the desired level we predict the risk of patients using the data provided.

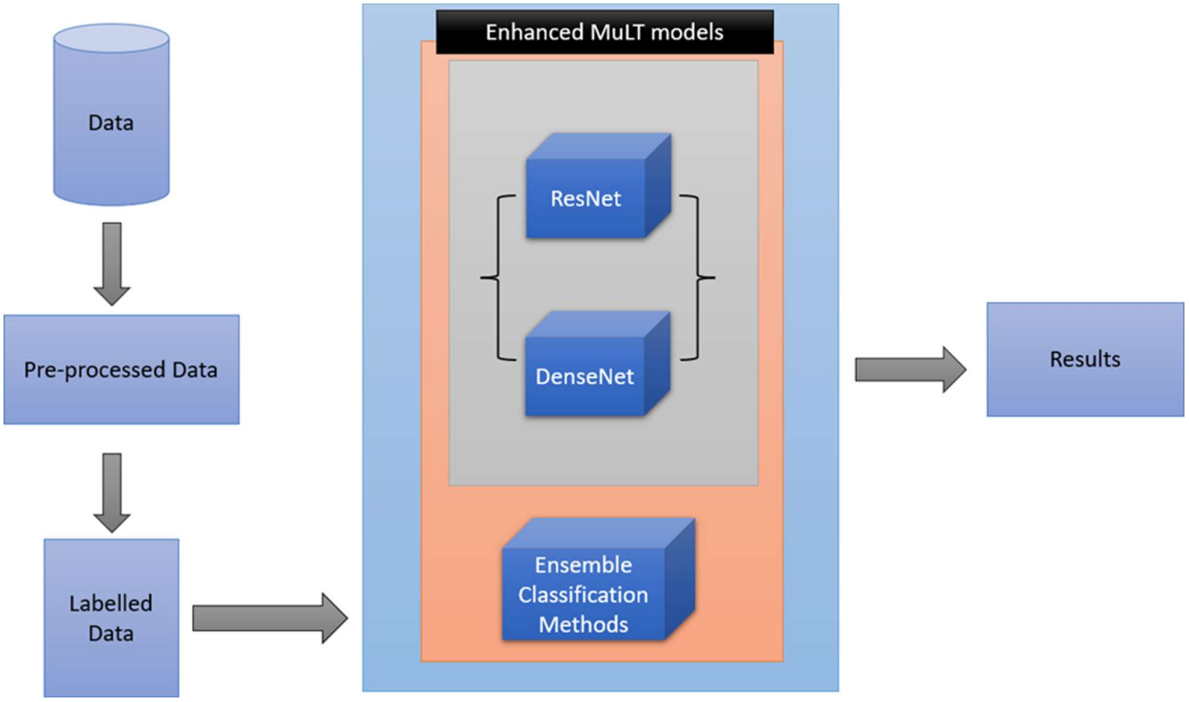

Figure 3: Workflow of Enhanced MuLT

The Enhanced MuLT algorithm proposed indicates multiple algorithms taken into consideration for analyzing the data provided. Figure 3 represents the overview of MuLT and the information regarding the process of execution.

To address a specific computational intelligence problem, many models, such as classifiers or experts, are strategically developed and merged in an ensemble learning process. Ensemble learning is largely used to enhance a model's performance in terms of classification, prediction, function approximation, etc. In this paper MuLT instances that are enhanced using ensemble learning are used to achieve the effective results.

A. DATA SET DESCRIPTION

The proposed work works on four different kinds of data that contains information of patients with various symptoms resembling cancer. The four different datasets are retrieved from public repositories. The Lung cancer, Pancreatic cancer, Breast cancer, Kidney cancer dataset contains multiple test samples with various symptoms or conditions. The data is a mixture of images and numerical values. The Lung Cancer dataset includes attributes like smoking, wheezing, chronic diseases, coughing etc., and the pancreatic cancer dataset has creatinine levels and the breast cancer takes into consideration the thickness, smoothness of the clumps for analysis and prediction. The kidney cancer dataset consists of training and validation set with around 4456 and 4440 instances each with normal and effected kidney images.

B. PRE-PROCESSING AND LABELLING OF DATA

The Pre- processing is a crucial phase in the data mining process, that can be defined as the altering or dropping of data before to usage in order to ensure or increase performance. In this paper the pre- processing involves steps like removing the null values, noisy data etc., Pre processed data helps in giving accurate results. Since semi supervised techniques are used, labelling of data is useful for better understanding of the system. Multiple algorithms and models are used for analysis which can be categorized into supervised and unsupervised techniques collectively called as semi supervised methods.

C. ENHANCED MuLT MODELS

We have developed the following subsystems of Enhanced MuLT implementing Ensemble Learning [9] for this project:

- **Logistic Regression**

The likelihood of a level is linked to a number of explanatory variables in the analytical modelling method of logistic regression. It is used to analyze datasets where the results are affected by one or more independent factors. The outcome is evaluated with a binary variable (in which there are only two possible results). It is used along with A type of analytical modelling known as logistic regression links a number of explanatory variables to the likelihood of a level. It is used to analyze datasets where the outcome is affected by one or more independent factors. [10] A binary variable is used to assess the result (in which there are only two possible results). To forecast a binary result (True/False, 1/0, Yes/No), a group of independent factors are used.

- **K Neighbour**

The K-nearest neighbour method is employed in the grouping and detection of patterns. It is frequently used in predictive analysis. When new data is provided, the K-NN algorithm selects the closest current data points. Any factors that can vary widely may have a major impact on the time between data points. The feature vectors and class labels are stored during the training phase. The data samples are ostensibly represented by K-NNs in a metric space. [11] The quantity is initially described in the classification phase using the K neighbours that are the most regular in the K training sample. The computer will then determine K of the most recent

neighbours of the sampled data.

- **Decision tree**

The supervised learning method known as a decision tree can be used to tackle classification and regression problems, but this strategy is widely used. It is a tree-structured classifier, where internal nodes stand in for the dataset's features, branches for the classification process, and leaves for the results of the classification. The Decision Node and the Leaf Node are the two nodes of a decision tree. While Leaf nodes are the results of decisions and do not have any

more branches, Decision nodes are used to create decisions and have numerous branches. Based on the characteristics of the supplied dataset, the test is executed or conclusions are made. It is a graphical tool for determining all reasonable solutions to a choice or problem within given parameters. [12] It is referred to as a decision tree since it starts with the root node and grows on succeeding branches to resemble a tree. [13] The CART method, which stands for Classification and Regression Tree Algorithm, is used to construct a tree. A decision tree simply poses a question and divides into subtrees based on the response (Yes/No).

- **Naïve Bayes**

The Naive Bayes algorithm, which is based on Bayes' theorem, is a supervised learning technique used to solve classification issues. Primary use cases include text categorization tasks that need a large and intricate training dataset.[14] Naive Bayes Classifier is one of the simplest and most successful Classification algorithms for rapidly building predictive machine learning models. A probabilistic classifier, like this one, bases its findings on the statistical probability that a certain item exists.

- **Random forest**

Random Forest is a kind of supervised learning that has found widespread use in the field of machine learning. It has several uses in the area of ML, [15] including classification and regression problems. This method relies on ensemble learning, in which numerous classifiers are combined to address a challenging issue and boost the model's accuracy and efficiency.

- **Support Vector Machine**

The Supervised Learning method known as Support Vector Machine (SVM) is widely used for both classification and regression problems. However, its primary use is in Machine Learning, [1 6] n amely for Classification problems.

Finding the optimal line or decision boundary that splits n-dimensional space into classes is the goal of the SVM method, making it easy to place fresh data points in the correct category in the future. The best limiting condition for choosing a choice is represented by a hyperplane.

- **Artificial Neural Networks**

One kind of learning algorithm is artificial neural networks (ANNs). The idea behind it was to simulate the behavior of real-world neural networks. Artificial neural networks (ANNs), which are computer models, are inspired by the brain networks of numerous animals. It can recognize trends and immediately pick up on new behaviours.

The popular machine learning technique of artificial neural networks (ANN) uses the past to predict the future, whereas the GA is an algorithm that may choose the most informative features to feed into ANN.

- **Adaboost**

AdaBoost, short for "Adaptive Boosting." is an Ensemble Method [17] that may be used in machine learning. AdaBoost is most often used to one-level decision trees, also known as single-split decision trees. These trees are also known as Decision Stumps.

- **Stacking**

Stacking Classifier at its most fundamental, stacking is an ensemble learning strategy in which the predictions of many classifiers (level-one classifiers) are utilized as fresh features to train a meta- classifier. You get to choose the classifier that serves as the meta-classifier.

- **Voting**

A Voting Classifier is a model in machine learning that takes into account the results of many other models to forecast an outcome (class).

The Classifier predicts the output class that is backed by an overwhelming majority of the input class by averaging the results of all classifiers fed into it. Instead of developing and evaluating separate models.[15][18] We may use them all to train a single model. which will then make output class predictions based on the models' average votes.

- **Bagging**

Bagging, also known an as Bootstrap ensemble learning aggregation. is approach that may be used to improve the efficiency and precision of machine learning algorithms. One of its applications is to cope with bias-variance trade-offs by decreasing the variance of a prediction model.

- INCEPTION is one of the best possibilities for RESNET V2 because it has been trained on more than a million images from the ImageNet collection, making it one of the convolutional neural networks. The network has 164 layers and can categorise images into 1000 different groups. In this network, we employ a technique called as skip connections. To connect layer activations to next layers, the skip connection skips over some intermediate levels. Consequently, a block is left over. ResNets are built by stacking these residual building blocks.

Instead of having layers learn the underlying mapping, the approach used by this network is to let the network fit the residual mapping. This type of skip link has the advantage of allowing regularization to bypass any layer that impairs architecture performance. As a result, disappearing or increasing gradients are not a problem while training an exceptionally deep neural network.

- For instance, DenseNet is a state-of-the-art neural network for visual object identification. DenseNet and ResNet are two popular deep learning networks, although they're not identical. ResNet's (+) additive strategy mixes the current layer's identity with the next layer, in contrast to DenseNet (-) concatenative approach (.) So, ResNet and DenseNet are the most commonly used approaches to deal with image data. A DenseNet is a particular sort of

convolutional neural network that makes use of dense connections between layers by using Dense Blocks, which link all layers directly with matching feature-map sizes. Each layer receives extra inputs from all earlier layers and transmits its own feature-maps to all later layers in order to maintain the feed-forward character of the system.

III. EXPERIMENT, RESULTS AND ANALYSIS

This paper includes four different datasets that were analyzed using multiple algorithms to predict the results. The experiments in this paper are performed out using 16GB RAM, an Intel Core i7, 11th generation CPU, and a Google Collaborator GPU with 4GB memory. The Lung Cancer dataset is processed under various algorithms and the depiction of the comparison of various models built is shown in figure 3.

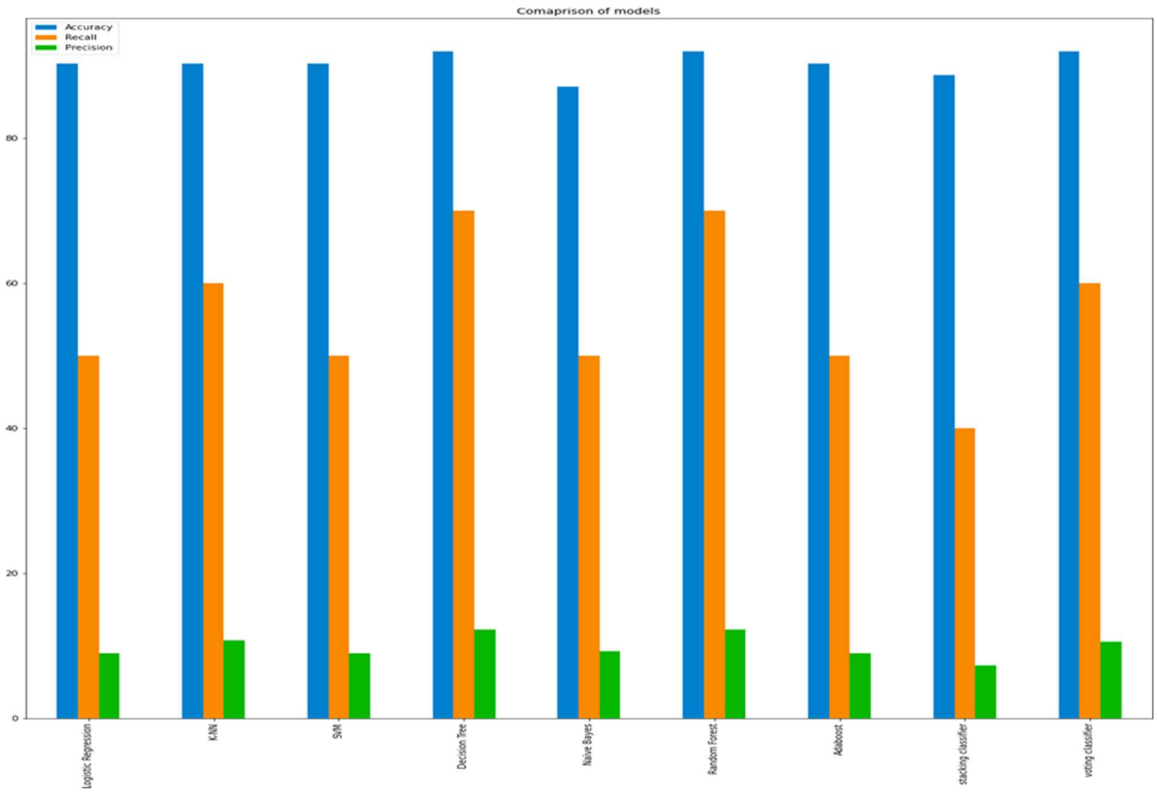

Figure .3: Comparison of Lung Cancer Models

The Figure .4 indicates the comparison of models built for prediction of pancreatic cancer when assessed using the datasets containing the creatinine levels

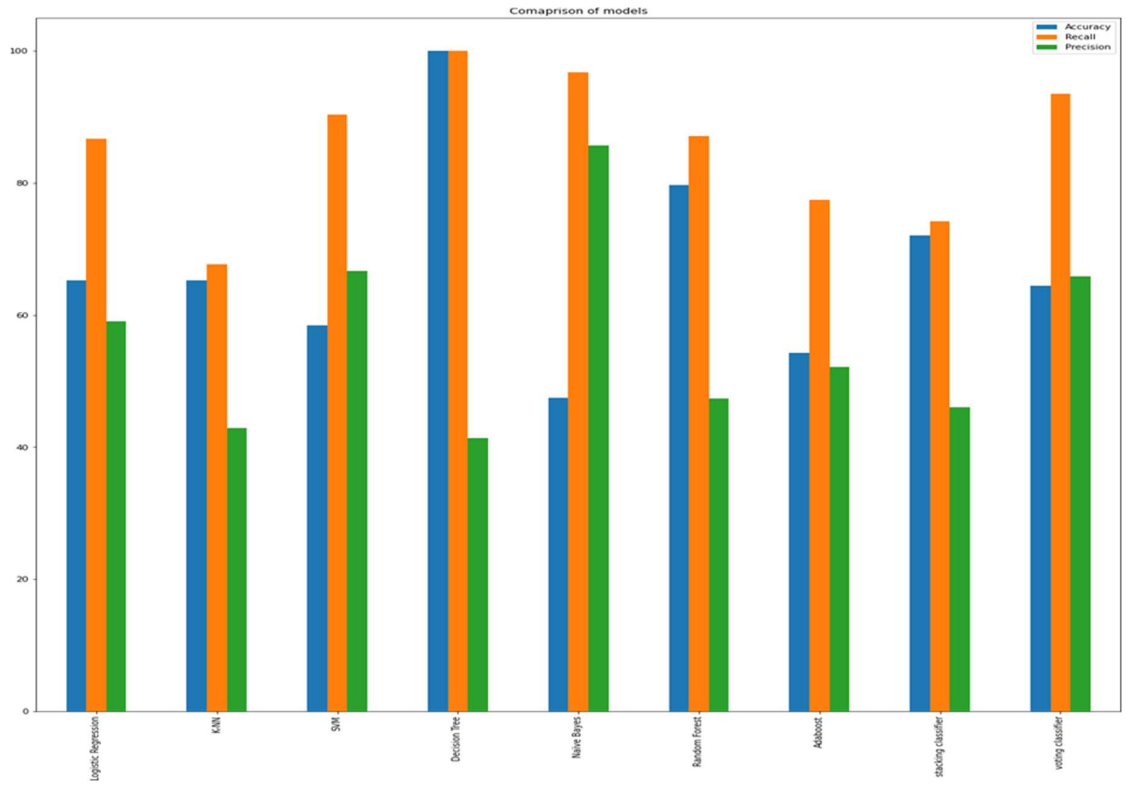

Figure .4: Comparision of models of Pancreatic Cancer

The results obtained are based on the analysis of various performance metrics. Following metrics were used in this work.

$$Accuracy = \frac{TP+FP}{TP+FP+TN+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F1\ Score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (4)$$

TP: True Positive FP: False Positive
TN: True Negative FN: False Negative

Equation (1), (2), (3) and (4) represents the formulae to calculate the accuracy, precision, recall and F1 Score respectively.

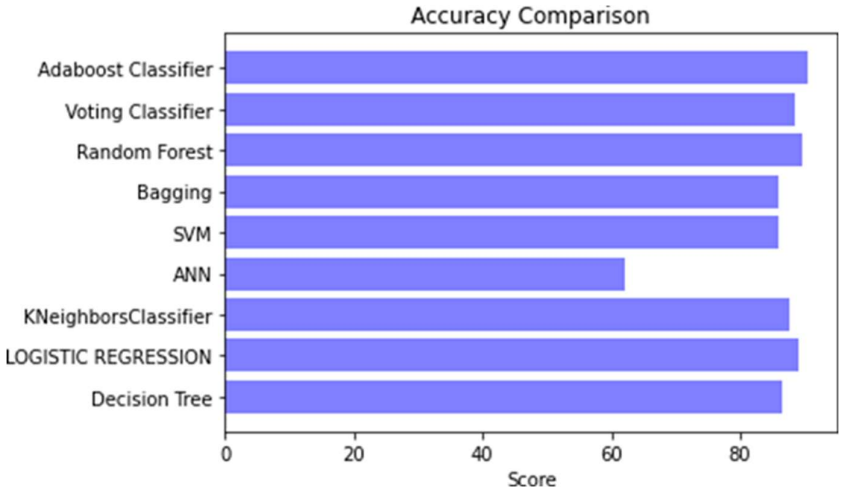

Figure. 5: Accuracy comparison of models of Breast Cancer

Figure 5 represents the comparison of accuracy obtained by the different models used in detecting the breast cancer and the one with highest accuracy is implemented in model building and to predict the results to be obtained.

The Kidney cancer uses two techniques ResNet and DenseNet for processing the image datasets. The accuracy obtained by both models is given in Table 1. The model with highest accuracy is implemented to predict the results of Kidney Cancer.

Table 1: Comparative analysis of models used in Kidney Cancer

Models	**Accuracy**
ResNet	93.48
DenseNet	89.80

CONCLUSION

As part of our investigations, we examined different publicly available data sets. The algorithms [17] outperformed the competitions in terms classification accuracy, which is particularly useful in situations when there is a high degree of missing data, a large degree of overlap across risk classes, and a scarcity of training data. Classification difficulty is reduced with respect to feature overlap because of the ability of the core modules of machine learning and deep learning algorithms to create new features from raw input.

Over the period of 11 months, our study gathered various datasets with multiple symptoms of cancer as the parameters for data analysis. These datasets were trained using multiple algorithms like K- nearest neighbor, Logistic regression, Support Vector Machines, Random Forest, Decision trees, Naïve Bayes, Artificial Neural Networks, Stacking, Bagging and Voting classifiers. Further the models like ResNet and DenseNet were applied to obtain the results with specific accuracy for each variant of cancer like, for Lung Cancer the accuracy achieved

is 93.54%, for Pancreatic cancer it is 93.40, for Breast Cancer it is 90.57% and for Kidney Cancer the accuracy obtained is 93.48%.

REFERENCES

[1] Essinger, Steve & Rosen, Gail. (2011). An introduction to machine learning for students in secondary education. 243 - 248. Doi:10.1109/DSP-SPE.2011.5739219.

[2] Reference: Mitchell, M. J., Jain, R. K and Langer, R. "Engineering and physical Sciences in oncology: Challenges and Opportunity Nature Rev. Cancer, volume es,"

17, issue II, pages 659-675, November 20 17, doi: 10.1038/nrc.2017.83.

[3] International Union for the Prevention of Cancer (2018). GLOB OCAN 2011 Online Updates Global Cancer Statistics from Previous Years. New global cancer statistics are available at https://www.uicc.org/new-global-cancer-data-globocan-2018.

[4] Designing a broad-spectrum integrative approach for cancer prevention and treatment," K. 1. Block et al., Seminars Cancer Biol., vol. 35, pp. 5276-5304, Dec. 2015, doi: 10.1016/j.semcancer.20 15.09.007.

[5] "SigneR: An empirical Bayesian approach to mutational signature discovery." Bioinformatics, volume 33, issue L pages 8- 16, January 201 7.14jR. A. Rosales, R. D. Drummond, R. Valiens, E. Dias-Neto, and 1. T. da Silva.

[6] The causes and consequences of genetic heterogeneity in the evolution of cancer, Nature, vol 50 L no. 7467, pp. 338-345, September 2013. 11RA Burrell, N McGranahan, J. Bartek, and C. Swanton.

[7] "Predicting treatment benefit in multiple myeloma through simulation of alternative treatment effects," by J Ubels, P

[8] Sonneveld, E. H. van Beers, A. Broul, M. H. van Vliet, and J. de Ridder, Nature Commun., vol. 9, no. L, pp. 1-10, Dec. 2018, doi: 10.1038/s41467-018-05348-5.

[9] Mai A. Shaaban, Yasser F. Hassan, Shawkat K. Guirguis , Deep Convolutional forest: a dynamic deep ensemble approach for spam detection in text ,2021 ,https://paperswithcode.com/paper/deep-convolutional-forest-a-dynamic-deep.

[10] Moffett Field, California, United States of America 17] Singularit y University. (2018). A Singularit y University Report on Rapidly Developing Healthcare Trends. I Online J. Available at: https://bit.ly/2Ti2ZIA

[11] Analysis of complexity indices classification problems: Cancer for gene expression data by A. C. Lorena, 1. G. Costa, N. Spolaor, and M. C. P. de Souto. Neurocomputing, vol. 75, no. L pp. 33-42, January 2012.

[12] Foundations of Machine Learning, 2nd edition. Cambridge, MA: MIT Press, 2018. LI Mohri. Mohri. Rostamizadeh and Talwalkar.

[13] In their article "Applications of machine learning in cancer prediction and prognosis," authors J.

A. Cruz and D. S. Wishart reference the January 2006 issue of Cancer Informal, volume 2, pages 1- 19.

[14] Rawal, Ramik. (2020). BREAST CANCER PREDICTION USING MACHINE LEARNING. 7. https://www.researchgate.net/publication/341508593_BREAST_CANCER_PREDICTION_ USING_ MACHINE_LEARNING

[15] Islam, M.M., Haque, M.R., Iqbal, H. et al. Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques. SN COMPUT. SCI. 1, 290 (2020). https://doi.org/10.1007/s42979-020-00305-w

[16] Ferjani, Marouane. (2020). Disease Prediction Using Machine Learning. 10.13140/RG.2.2.18279.47521. https://www.researchgate.net/publication/347381005_Disease_Prediction_Using_Machin e_Learning

[17] D. Yifan, L. Jialin and F. Boxi, "Forecast Model of Breast Cancer Diagnosis Based on RF-AdaBoost," 2021 International Conference on Communications, Information System and Computer Engineering (CISCE), 2021, pp. 716-719, doi: 10.1109/CISCE52179.2021.9445847.

[18] F.J. Shaikh, D.S. Rao, Prediction of Cancer Disease using Machine learning Approach, Materials Today: Proceedings, ISSN 2214-7853, https://doi.org/10.1016/j.matpr.2021.03.625. (https://www.sciencedirect.com/science/article/pii/S2214785321027206)