International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 2s
DOI: https://doi.org/10.17762/ijritcc.v10i2s.5918
***Article Received**: 08 October 2022 **Revised**: 07 November 2022 **Accepted**: 26 December 2022*

Statistical Analysis and Deep Learning Associated Modeling for Early stage Detection of Carinoma

[1]Dr. K. Rangaswamy, [2]D. Dhanya, [3]B. Rupa Devi, [4]Sateesh Kumar Reddy C, [5]R. Obulakonda Reddy
[1]*Associate Professor*, Department of CSE, Rajeev Gandhi Memorial college of Engineering and Technology, Nandyal, Andhra Pradesh.
Email: rangaswamy19@gmail.com
[2]*Assistant Professor*, School of Engineering, CSE, Presidency University, Bangalore, Karnataka.
Email: dornadhula.dhanya@presidencyuniversity.in
[3]*Associate Professor*, Department of Computer Science & Engineering, Annamacharya Institute of Technology and Sciences, Tirupati.
Email: rupadevi.aitt@annamacharyagroup.org
[4]*Assistant Professor, Department of Electronics and Communication Engineering, Narayana Engineering College, Nellore. Email: satishreddic@gmail.com*
[5]*Associate Professor, Department of Computer Science and Engineering (Cyber Security), Institute of Aeronautical Engineering, Dundigal, Hyderabad. Email: rkondareddy@gmail.com*

Abstract—The high death rate and overall complexity of the cancer epidemic is a global health crisis. Progress in cancer prediction based on gene expression has increased in light of the speedy advancement using modern high-throughput sequencing methods and a wide range of machine learning techniques, bringing insights into efficient and precise treatment decision-making. Therefore, it is of significant interest to create machine learning systems that accurately identify cancer patients and healthy people. Although several classification systems have been applied to cancer prediction, no single strategy has proven superior. This research shows how to apply deep learning to an optimization method that uses numerous machine learning models. Statistical analysis has helped us choose informative genes, and we've been feeding those to five different categorization models. The results from the five different classifiers are ensembled in the next step using a deep learning technique. The three most common types of adenocarcinoma are those of the lungs, stomach, and breasts. The suggested deep learning-based inter-ensembles model was tested with deep learning-based algorithms on Carcinoma data. The results of the tests show that relative to using only one set of classifiers or the simple consensus algorithm, it improves the precision of cancer prognosis in every analyzed carcinoma dataset. The suggested deep learning-based inter-ensemble approach is demonstrated to be reliable and efficient for cancer diagnosis by entirely using diverse classifiers.

Keywords-Statistical Analysis; Parametric Analysis; Data Testing; Cancer; Deep Learning methods.

I. INTRODUCTION

Cancer is a group of disorders characterized by uncontrolled cell development with the ability to invade neighboring tissues and metastasize. The GLOBOCAN project estimates that there were 18.5 million deaths annually worldwide in 2021 (excluding skin cancers other than melanoma), accounting for approximately 18.6% of all deaths from cancer that year. The ability to detect and diagnose cancer early is crucial for its treatment because cancer is a leading cause of death and suffering. Cancer research has had a steady development throughout the past few decades. The use of gene expression levels is one of the many active study areas in cancer prediction. Cancer treatment and detection have benefited greatly from gene expression data analysis. An important and pressing problem for doctors today is the development of more reliable methods for cancer prognosis [1]–[3]. Recent years have seen a surge in the prevalence of computer-aided approaches; as a result, machine learning methods have been applied to cancer detection, with researchers regularly exploring new prediction algorithms. Data from Egypt's National Cancer Registry Program was used in a study comparing three classifiers (support vector machines (SVMs), k-nearest neighbors (kNN's), and Naive Bayes) for feature selection and classification (N.B.'s). We found that in classification accuracy, SVMs with polynomial kernel functions outperformed kNN and N.B.s. They included a thorough evaluation of SVMs against random forests for cancer diagnosis. It was found that support vector machines (SVMs) outperformed random forests (R.F.s) outperformed Support Vector Machines (SVMs) in nine data sets, while the results were comparable in three other datasets. To get these outcomes, the entire gene set was used. The gene selection approach yielded comparable outcomes. As seen from the vast literature on cancer prediction, none of the machine learning techniques are without flaws, and they may all be inadequate in some way or another during the categorization process. For example, SVMs have difficulty determining a proper kernel function, and R.F.s may bias classification results toward the group with more samples while solving the over-fitting problem that plagued decision trees (D.T.s). Since each machine learning approach has the potential to outperform others or have flaws in specific scenarios, it stands

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 2s
DOI: https://doi.org/10.17762/ijritcc.v10i2s.5918
***Article Received**: 08 October 2022 **Revised**: 07 November 2022 **Accepted**: 26 December 2022*

to reason that a strategy that leverages the strengths of various machine learning approaches will result in better overall performance. To improve the reliability of predictions, various research has been discussed in the literature that combines different models. For instance, they came up with Bagging, which takes the results of decision trees built from different randomly chosen parts of the training examples and averages them together to reach a final judgment. With the advent of Boosting, it is now possible to combine the classification outputs using weighted votes, which are updated after each training cycle and based on the importance of each training sample. They proposed using linear regression to integrate neural network outputs, a technique that would become known as Bagging and Boosting and be used for cancer clmassification using microarray data. Using three standard cancer data sets, they combined the results of four classifiers using the majority voting algorithm. Different types of machine learning are used for The Stacking and the majority vote [20]. The majority voting technique is the most popular method for combining classifiers in classification problems, but it is still insufficiently simplistic to uncover detailed information. Stacking is a more potent ensemble strategy since it incorporates a learning method into the combination phase. Deep learning has emerged as a powerful learning method with various benefits because of the limited number of research that has demonstrated its effectiveness in biomedicine. In contrast to popular vote, which only considers concatenation among classifiers and requires human intervention, deep learning has the potential to "learn" the complex structures, especially nonlinear structures, of the original enormous data sets without any human intervention. With this in mind, we employ deep learning in the Stacking-based evolutionary algorithms of several classifiers to clarify their mysterious connections further. This paper uses deep neural networks to ensemble five classification models for cancer prediction: kNN, SVMs, DTs, R.F.s, gradient-boosting decision trees (GBDTs), and tumor states. We apply the gene expression differential analysis to select relevant and informative genes while avoiding over-fitting. After that, the chosen genes are fed into the five different classifiers [21]. The prediction result is then obtained using a deep neural network to ensemble the results from the five classification models. To test the efficacy of the suggested technique, we use publicly available data from lung, stomach, and breast tissues. According to the final findings, the suggested deep learning-based multi-model clustering algorithm delivers a better accurate estimate than classification models or the majority voting algorithm while also making better use of the limited clinical data.

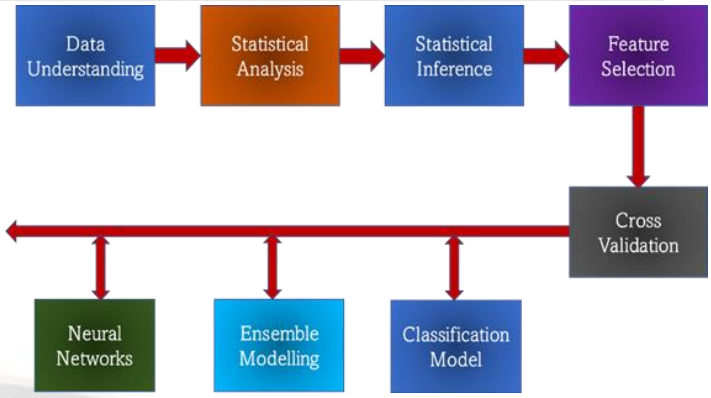

Figure 1: Functional Structure of the research process.

Fig. 1 is a flowchart depicting the suggested statistical and deep learning-based ensemble strategy. Starting with differential expression analysis, the most informative features, i.e., substantially differentially expressed genes, are selected and supplied in the subsequent classification step. Next, we use S-fold cross-validation to split the raw data into S sets for training and testing. Next, multiple classifiers are learned from the training sets, each consisting of S 1 of the S groups, and then used in the matching test set, which is the remaining group of the S groups, to produce the classification model of the samples. Finally, to reduce the generalization error and obtain a more precise result, we employ a deep neural network classifier to aggregate the predictions from the first stage [4]- [6].

II. DATA ANALYSIS USING STATISTICAL MODELING WITH SAS-JMP TOOL - METHODS

In the dataset, there are various parameters. To understand the data and its interpretation, the SAS JMP tool was used for statistical modeling. The carcinoma dataset consists of different information and related parameters. A few essential parameters are shown in the figures of distribution. Tumor information with the distribution.

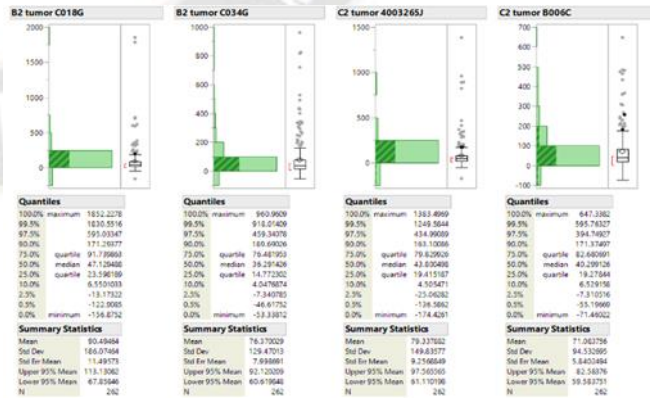

Figure 2: Tumor B2, C2 category – Statistics

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 2s
DOI: https://doi.org/10.17762/ijritcc.v10i2s.5918
***Article Received**: 08 October 2022 **Revised**: 07 November 2022 **Accepted**: 26 December 2022*

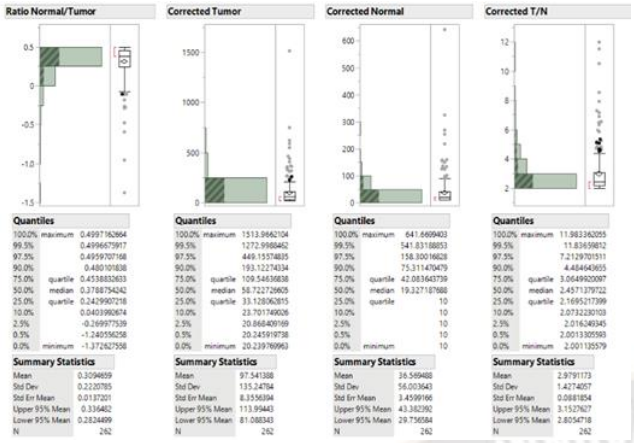

Figure 3: Tumor, Normal Distributed statistics.

As more variables and data are used in statistical analysis data, it becomes more challenging to create classification models. Over-fitting and deterioration of classification ability are more likely to occur in clinical practice due to the small number of available cancer samples relative to the number of characteristics. When faced with such difficulties, selecting features is a valuable tool. The difficulties caused by a short sample and a high data complexity can be alleviated by restricting the subspace to a smaller subset of features before training a classification model. In this work, we use the Distribution technique to identify candidate genes that aid the subsequent classification. The data distribution modeling approach is commonly employed to determine if a change in read count observed for a specific gene is statistically significant (i.e., more extraordinary than that which would be predicted owing to natural random fluctuation). With the help of a BH-adjusted p-value and a fold change threshold, significantly differentially expressed proteins can be filtered out in a differential expression study.

Main Objectives Over existing Methodologies under Cancer Detection models. The critical finding significantly shows the differences between existing and proposed hybrid models like Ensemble methods and Neural network-based Deep learning methods.

1. In Detail, Analysis was carried out on the Carcinoma Dataset to understand the patterns in the data.
2. Statistical Inference, key findings calculated to show the differences between Normal conditions and Tumor Conditions
3. Analyzed the Tumor to Normal and Normal Tumor conditions to predict cancer from the Normal and Tumor states using Statistical, Machine learning based ensemble, and Deep learning / Neural network based modeling.

A. Cross Validation – Data Modeling

Numerous parameters may determine the complexity of several different classification models. The goal is to determine the optimal values for the complex variables that will yield the highest predictive accuracy on new data for a given application. If data are abundant, one can select a model with minimal effort by splitting the data into a training set, a validation set, and a test set. Various models are taught on the training dataset, evaluated and chosen on the testing set, and assessed on the testing set. The best-performing complicated model is chosen among those trained; this model is effective by the validation set. However, data availability to train and test is constrained in a real-world setting, which increases the generalization error. Cross-validation is a method for minimizing the generalization error and avoiding over-fitting. Figures 2 and 3 depict data distribution concerning the S-fold cross-validation method used in this paper with S equal to 4. Using S-fold cross-validation, we divide the full dataset D into S independent subsets, D1, D2, and D.S., with the same data distribution across all groups. After that, we split the remaining population into the test set and the S-1 groups used for training.

The process was repeated for each of the S - 1 possible combinations of groups, and the average performance rating overall S iterations were calculated. To prevent over-fitting, we employ S-fold cross-validation to independently perform a model selection for each classifier and produce new sets of data for the ensemble stage.

Regarding the Carcinoma dataset, many features are represented with various functionalities. Mainly it showed the tumor and normal conditions concerning ratios. Fig 4 shows the tumor vs. normal conditions. Here, along with various other parameters, are also mentioned in detail. Fig 4 shows the Average tumor vs. Average normal conditions, displayed using statistical analysis points. In that scenario, it observed that few factors significantly show that there are changes to be turned into cancer [7]–[9]. In alignment with the objective of this research is to detect cancer from tumor-level chances.

III. RESULTS AND DISCUSSIONS

A. Ensemble Learning Algorithms – Modeling

Here, we evaluate the forecast performance of five well-known classification algorithms for separating healthy tissue from cancerous tissue, following preprocessing the data sets. When it comes to the initial step of the classification process, we use k-nearest-neighbor (kNN), support vector machines (SVMs), decision trees (D.T.s), random forests (R.F.s), and gradient boosting decision trees (GBDTs). The five categorization approaches discussed below are all highly accurate in real-world situations, and their merits are discussed in detail. When minimal information is available about the data's distribution, k-nearest neighbors (kNN) can be an effective categorization method. In

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 2s
DOI: https://doi.org/10.17762/ijritcc.v10i2s.5918
***Article Received**: 08 October 2022 **Revised**: 07 November 2022 **Accepted**: 26 December 2022*

k-nearest neighbor classifiers, the distances between data are calculated by mapping them to a metric space.

Classifying a test sample according to the most prevalent class in its k-nearest training samples is the goal of k-nearest neighbors, and this is accomplished by computing the distance measure between a test sample and the training samples. For SVMs, the first step is to transfer the input sequence into a higher-dimensional feature space and then find a hyperplane that divides the data into two classes. The chasm between the two groups is maximal. Then, new samples are projected into the same region, and their predicted category is determined by which side of the divide they land on with greater certainty. Decision trees (D.T.'s) have the form of trees, with the nodes standing in for the input parameters and the leaves representing the decisions made. Due to the unique structure, we can accurately forecast the data category as we travel down the tree to classify a new sample. Cancer prediction using R.F.s is a relatively new field.

Combining tree predictors using the same distributional random vector, R.F.s is an ensemble learning method. Ultimately, the best model is produced by the class that obtains the most votes from individual trees in the forest. Combining multiple decision trees into one robust model for prediction is the goal of GBDTs, a machine-learning technique. GBDTs implement a generalization by permitting the optimization of an arbitrarily differentiable loss function, and they construct the model in a stage-by-stage approach like previous boosting methods. Six different approaches are presented: three traditional ones (kNN, SVMs, and D.T.s) and two more modern ones (R.F.s and GBDTs). There is some evidence in the literature to imply that kNN is one of the easiest classification methods, particularly for data with an uncertain distribution. However, the performance of a classifier can be significantly impacted by the choice of the number k, and kNN is susceptible to redundant data and requires effective feature extraction before classification. It's safe to say that SVMs are the most popular and successful algorithm for identifying cancer types [10]–[12]. SVMs face a difficult task, however: determining which kernel is best suited to solve a given problem.

Nonlinear instances do not have a general solution, making it impossible to ensure the correctness of the predictions. D.T.s are the most basic and extensively used classification approach across many different domains; nevertheless, they are generally ineffective when differentiating between normal and cancerous samples. The latter two techniques, R.F.s and GBDTs, are ensembles of D.T.s that evolve to solve the over-fitting issue; nonetheless, the classification result may be biased toward the group with more samples.

TABLE 1- K.N.N. Algorithm Analysis

K	Count	R-Square	RASE	SSE
1	262	-0.9649	0.08264	1.78921
2	262	-0.483	0.07179	1.3504
3	262	-0.3494	0.06848	1.22875
4	262	-0.1381	0.06289	1.03634
5	262	-0.0383	0.06007	0.9455
6	262	-0.0054	0.05911	0.91547
7	262	0.00158	0.05891	0.90915
8	262	-0.0012	0.05899	0.91166
9	262	-0.0047	0.05909	0.91489
10	262	0.00368	0.05885	0.90724

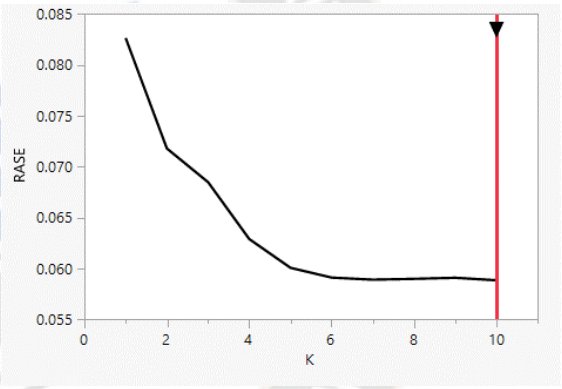

Figure 5: K N.N. Algorithm Analysis K - Fitting.

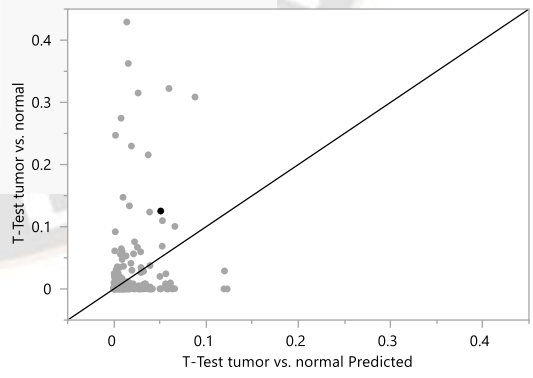

Figure 6: K N.N. Algorithm Analysis T- Test Tumor Vs. Normal condition

Keeping in mind that different approaches have different drawbacks, we develop an ensemble strategy to leverage the benefits of several approaches while avoiding their drawbacks. In this case, we choose both basic and evolutionary approaches to improve the robustness of our ensemble classifier.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 2s
DOI: https://doi.org/10.17762/ijritcc.v10i2s.5918
***Article Received**: 08 October 2022 **Revised**: 07 November 2022 **Accepted**: 26 December 2022*

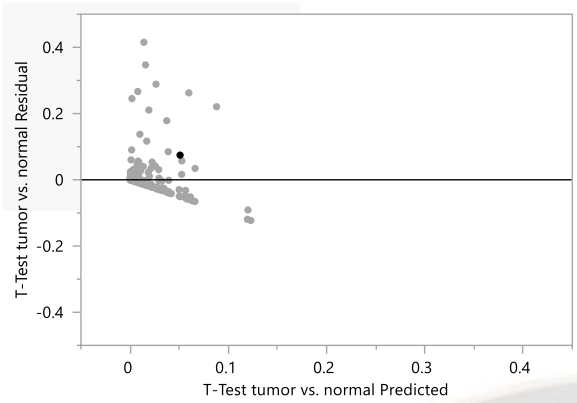

Figure 7: K N.N. Algorithm Analysis T- Test Tumor Vs. Normal condition

Table 1 shows the values for K – Nearest neighbors algorithms fitting models. R-Square, RASE, and S.S.E. values represent the K.N.N. algorithm's function fitment. Fig 5,6 and 7 show the kNN algorithm's training and testing scenarios. Similarly analyzed with Support Vector Machines algorithm for understanding the fitment analysis for cancer prediction from the level of tumor and normal conditions shown in Fig 8 and Fig 9.

TABLE 2- SVM Algorithm Analysis

Response	T-Test tumor vs. normal
Validation Method	**KFold**
Kernel Function	Radial Basis Function

Measure	Training	Validation
Number of rows	210	52
Sum of Frequencies	210	52
RASE	0.0580823	0.0466788
R-Square	0.0973591	0.0950323
Number of Support Vectors	89	89

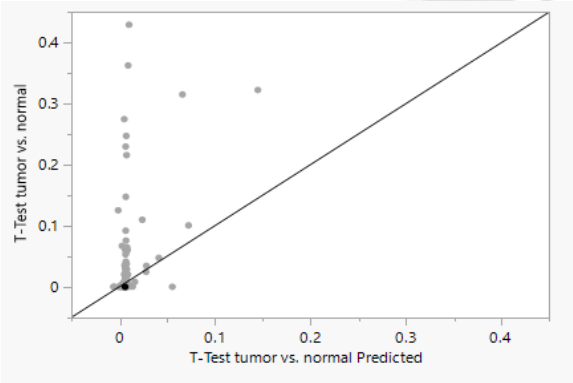

Figure 8: SVM Algorithm Analysis T- Test Tumor Vs. Normal training set

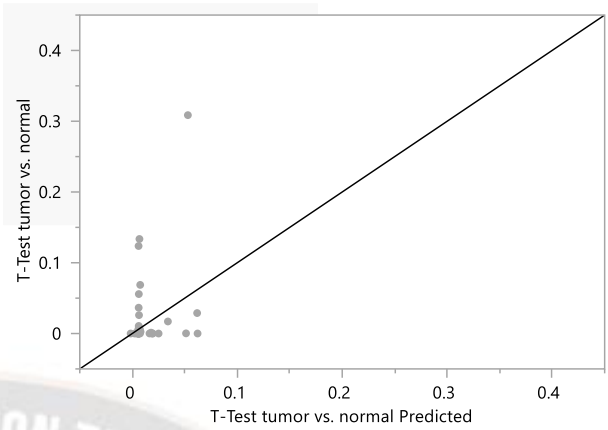

Figure 9: SVM Algorithm Analysis T- Test Tumor Vs. Normal testing set

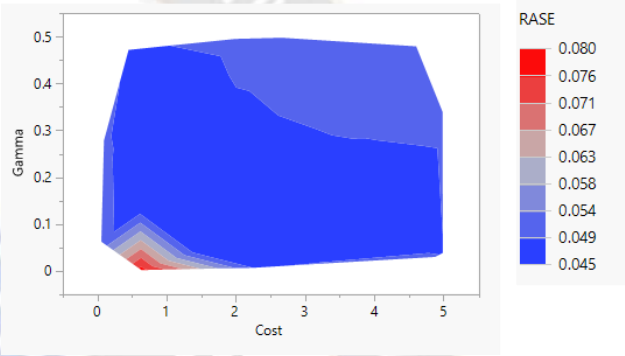

Figure 10: SVM Algorithm Analysis T- Test Tumor Vs. Normal modeling accuracy -

Gamma – 0.07779 and Cost – 3.03 analysis shows the levels of the tumor vs. normal conditions to cancer-turning conditions concerning the carcinoma dataset.

B. Deep Learning Based Modeling and Analysis

First, there are several cancer prediction categorization models to choose from, but none are perfect and may be inaccurate in some vital respect. When used together, various categorization algorithms may outperform their solo versions. With a multi-model ensemble, a group of models' predictions is used as inputs to a more advanced learning model. The second-stage model training aims to provide an optimal set of predictions by combining the predictions from the first-stage models. This paper uses deep learning as an ensemble model to combine the results of several different classifiers into a single precise estimate. Neural networks, which take their cues from the brain's structure and operation, find widespread use in various contexts. It is possible to train a neural network to produce an output from several inputs. It can be trained to approximate nonlinear functions given a set of characteristics and an end goal, with one or even more nonlinear layers (hidden layers) inserted between both the output and input layers[8], [9], [13]–[16]. To learn complicated patterns from high-dimensional

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 2s
DOI: https://doi.org/10.17762/ijritcc.v10i2s.5918
***Article Received**: 08 October 2022 **Revised**: 07 November 2022 **Accepted**: 26 December 2022*

raw data with minimal supervision, deep learning employs deep neural networks having multiple hierarchical hidden units of nonlinear processing information. An example neural network is presented. Input neurons are the neurons found in the leftmost layer, which is called the input layer. The output layer is the rightmost and contains an output neuron. The buried neurons make up the layers in the middle. We create an objective function that calculates the deviation from the actual scores to the anticipated scores to ensure accurate sample classification. Then, the machine learns from the training samples and adjusts the values of the variables that describe the input-output function internally, leading to a minor inaccuracy.

The stochastic gradient descent (S.G.D.) algorithm is typically utilized for this machine learning method. Layer L1 is the input layer, and layer Lnl is the output in a deep neural network where the number of levels is represented by nl and Ll represents each layer. Similarly, we'll refer to the total number of neurons in layer l as sl. Parameters for the neural network are written as W = [W1, W2,..., Wnl] and b = [b1, b2,..., bnl], where Wlij, j = 1, 2,..., sl1, I = 1, 2,..., sl, l = 2, 3,... To illustrate how the S.G.D. can be used to train a neural network, let's imagine that we have m samples of data labeled "(x 1, y1),(x 2, y2),...,(xm, ym)" to use as a training set. For this discussion, let's refer to the cost function (objective function), where is a weight decay parameter that controls the relative importance of a mean-squared error term and a regulation term that limits the orders of magnitude of the weights to prevent over-fitting. A neural network's nonlinear hypothesis, hW,b(x), is written as

The rectified linear unit (ReLU) f(z) = max0, z has been the most widely utilized nonlinear function in this context in recent years [18]. The ReLU typically learns faster in inter-deep neural networks than traditional nonlinear and logistic sigmoid functions. I denote the engagement of unit I in layer l, and the sum is denoted by z l I for a given sample.

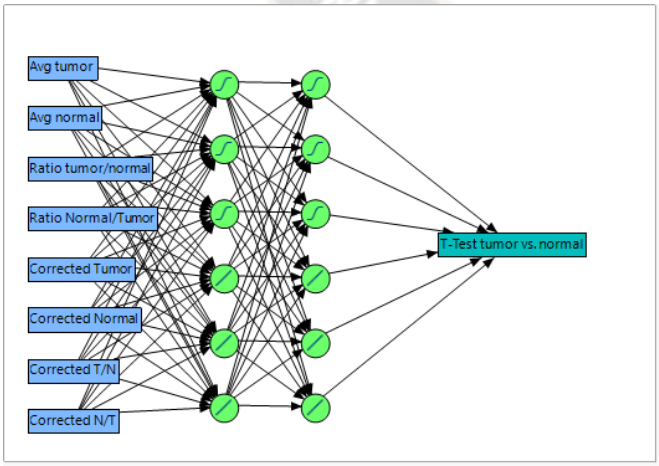

Figure 11: Neural Network Algorithm Analysis on T- Test Tumor Vs. Normal data with Ratios and Corrected Tumors to Normal tumors on Cancer prediction.

Calculating an individual unit's activation is known as forward propagation. The objective of S.G.D. is to minimize J (W, b) by tuning the values of W and b. In the first step of S.G.D., we set each Wij and bli to a small random value close to zero, then modify each loop's parameters. This is the rate at which one picks up new information. Then, we use the backpropagation procedure to calculate the partial derivatives. Assuming we have training examples (x, y), we can describe the backpropagation technique in more depth as follows. The first step is to perform forward propagating calculations to determine the activity of each unit in layer L2 from the input layer Lnl to the final output layer Lnl. Next, determine the residual for every unit I in layer nl. Determine the residual for layer l, l = nl 1, nl 2,..., 2, for each I in the layer.

Then, get the needed partial derivatives by calculating them. The goal of backpropagation is to derive anything from its antecedents. To train a neural network, the function J(W, b) must be minimized as often as possible by iterations of the S.G.D. The research presented below proposes a 5-fold stacked multi-model ensemble approach based on deep learning. A diagrammatic representation of the entire procedure is presented in Fig. 4. The first step is to partition the original data set D into five smaller subsets, D1, D2, and D5, each of which contains labeled points drawn independently and identically dispersed based on the same distribution (i.e., Di = xi, yi, I = 1, 2,..., 5). Initially, we choose D2 + D3 + D4 + D5 as our training set and D1 = x1, y1 as our test set. Specifically, given input x1, the five proposed classification hypotheses are h1(x1), h2(x1), and h5(x1), where hi(xi) is a binary variable, and the subscript denotes the ith model I of hi(x1).

Following the first-round classifications, we combine the predictions from each model into one large set, H1 = [h1(x1), h2(x1), h5(x1)], and then append the associated label y1 to this to create a new data set, D'1, for use in the second phase. Following the S-fold cross-validation strategy detailed in the Methods section, this process is done five times. Five additional data sets are obtained as a result: D'1, D'2, D'5, where Di = Hi, Yi, I = 1, 2, 5. We use deep neural networks as the ensemble classifier to further refine our results in the second phase. We employ a five-layer neural network to distinguish between healthy and cancerous tissue samples. Five neurons make up the network's input layer, and they stand in for the sample's characteristics in the new data set. We test various configurations of the number of nodes per layer in the hidden representation to enhance classification accuracy. One neuron, whose output was either 0 (standard) or 1 (tumor), is in the network's output layer. We use 5-fold cross-validation and average the results to reach this point [17], [18].

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 2s
DOI: https://doi.org/10.17762/ijritcc.v10i2s.5918
***Article Received**: 08 October 2022 **Revised**: 07 November 2022 **Accepted**: 26 December 2022*

TABLE 3- Multilayer Neural Network Algorithm Analysis

Measures	Value
R-Square	0.0202103
RASE	0.0584348
Mean Abs Dev	0.0192947
-LogLikelihood	-392.3317
SSE	0.5941446
Sum Freq	174

The deep learning-based ensemble method "learns" the associations automatically, as opposed to the more traditional weight average and majority vote algorithms in standard ensemble strategy, which only evaluate the linear correlations across classifiers and require operator participation. It is impossible to forecast with any certainty using an essentially linear relation since the connections between the many classifiers and the labeling of samples are unknown. Nonetheless, our approach's second stage employs deep learning, which automatically learns complex relationships (especially nonlinear ones) and necessitates minimal engineering by hand. To ensure accurate predictions, the deep learning-based inter-ensemble method uses all available data.

All phases of cancer are included in these data sets, which were compiled from a wide range of clinical conditions, ages, sexes, and ethnicities. For this profile, we looked at tumor samples collected from individuals who had never had chemotherapy or radiation therapy. Table 1 displays detailed information about the datasets. Both raw data counting and normalized fragments per kilobase per million (FPKM) data were employed in our method. The substantially differentially expressed genes were chosen from the raw count data, and the normalized FPKM values were used in the formal recognition and ensemble procedure.

This process analyzed various parameters based on the modeling with respected Corrected Tumor to Normal condition and Normal to Tumor conditions to find cancer-related attributes using Neural Networks based Deep learning models.

TABLE 4- Difference: Corrected T/N-T-Test tumor vs. normal

Corrected T/N	2.97912	t-Ratio	33.9914
T-Test tumor vs. normal	0.01958	DF	261
Mean Difference	2.95954	Prob > |t|	<.0001*
Std Error	0.08707	Prob > t	<.0001*
Upper 95%	3.13098	Prob < t	1.0000
Lower 95%	2.78809	-	-
N	262	-	-
Correlation	0.32519	-	-

We used the 5-fold cross-validation approach to average the predictions from the five classification methods we used in the first stage: k-nearest neighbor, support vector machines, decision trees, random forests, and gradient-boosting decision trees. After that, we used a multi-model ensemble approach with a deep neural network to combine all the predictions from the first stage. Because of the larger sample size and the lower dimension of the new data set, thanks to 5-fold cross-validation, deep neural networks can be used, and the prediction accuracy is improved.

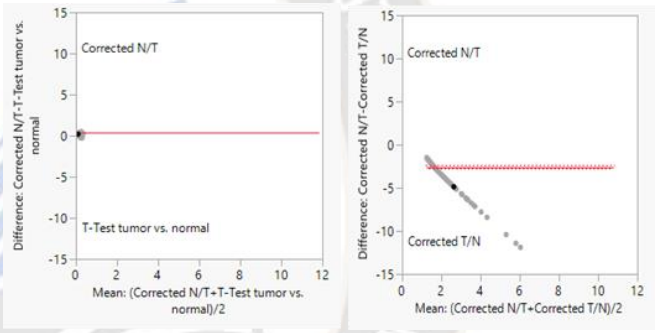

Figure 12: Finding the relation concerning Normal / Tumor – Corrected conditions.

Three datasets (LUAD, STAD, and BRCA) are used to assess the predictive reliability of our proposed ensemble technique, as well as individual methods, majority voting, and individual methods. To evaluate the efficacy of various techniques, the R.O.C. curve is often employed. The receiver operating characteristic (R.O.C.) curve is a graph in statistics that shows how well a binary classifier does when the threshold for making a distinction between classes changes. It is generated by comparing the true positive rate (the proportion of actually correct positives) to the false positive rate (the proportion of false negatives) for a range of confidence levels. When comparing models, it is common to practice estimating the area under the curve (A.U.C.), which represents the likelihood that a classifier will place a randomly selected good example higher than a randomly selected negative one. Precision-recall (P.R.) curves are necessary for handling the highly skewed data we utilized.

The area under a P.R. curve is commonly used to evaluate the connection between the recall and precision and the effectiveness of a classifier. If the area is large, then the precision and recall are good. Table 3 displays the findings of the predicted

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 2s
DOI: https://doi.org/10.17762/ijritcc.v10i2s.5918
***Article Received**: 08 October 2022 **Revised**: 07 November 2022 **Accepted**: 26 December 2022*

accuracy analysis. For each cancer dataset, we compare our proposed ensemble method's recognition rate to that of five classifiers and the simple majority technique, as shown in the table below. As seen in the table, integrating many models yields far more reliable performance than utilizing a single model for categorization. More so than majority voting, our deep learning-based multi-model ensemble approach improves upon predictions with an accuracy of 99.20%, 98.78%, and 98.41% for the LUAD, STAD, and BRCA datasets, respectively. Fig 11 displays the Neural Networks structure for the cancer case datasets.

The figures show that by merging several different classifiers, the integration technique improves classification performance over the best performance of any individual classifier. The suggested deep learning-based ensemble technique for all three datasets outperforms majority voting thanks to its inherent capacity to automatically learn and find latent structure. To illustrate, Fig. 12 displays the three related P.R. curves. As seen in the diagram, the suggested ensemble method outperforms individual classifiers and majority voting, obtaining a region that is either greater than or equal to theirs. In addition, we find that the suggested ensemble method is effective when dealing with skewed statistics, which reflects the inequity of clinical samples.

Results show that the suggested deep learning-based inter-ensemble method outperforms the individual classifier and the simple majority method in cancer prediction. Timely and correct diagnosis is crucial because of cancer's intricacy and high mortality rate. Thus, employing computer-aided approaches to increase the accuracy of predictions is very helpful in cancer treatment. We compared our suggested multi-model ensemble technique and five individual classification models. Traditional and cutting-edge classifiers have been used extensively in cancer forecasting, and those five methods make up the bulk of the list. An earlier study found that SVMs and R.F.s can be effective classifiers but that one may be superior to the other depending on the data set being analyzed. This suggests that alternative classifiers may experience the same problem, demonstrating that every approach has limitations. This realization drove us to propose an approach for merging many classifiers into a single framework, which we believe will result in a more robust and objective model for classification. Based on our analysis of three datasets, we can conclude that the inter-ensemble method improves accuracy over the individual performance of five classifiers.

Furthermore, the R.O.C. curves demonstrate that the performance of a single classifier's predictions is inconsistent across datasets. This is likely the result of classifiers' different sensitivities to various forms of input, as well as to small or large sample sizes and too redundant features. On the other hand, our suggested technique continues to train each classifier's strength even while it ensembles the results from the five classifiers. Higher-accuracy classifiers play a more significant role, while lower-accuracy classifiers' interference information is filtered out in this method. Since the benefits of each classifier are taken into account and used to their full potential, improved prediction performance is achieved. The proposed ensemble technique based on deep learning was put through additional tests, and its prediction accuracy was compared to that of the simple majority algorithm. As with many other applications of ensemble methods, the simple majority algorithm is used in cancer prediction. Our findings corroborate those who found that majority voting outperformed SVMs and kNN for classifying cancer data.

Additionally, our deep learning-based ensemble method outperforms the majority voting algorithm in terms of accuracy and area under the curve (A.U.C.). The simple majority algorithm only takes into account linear relationships, so this may be why the outcomes are what they are. Our proposed method employs deep learning in an ensemble setting to dynamically learn hidden complex structures, such as nonlinear ones, instead of the majority voting algorithm. The output of various classifiers and the connections between them are considered through deep learning training, which reveals and best fits the unknown connections among the classifier and the labeling of samples. And so we get better cancer forecasts. Our findings also show that the ability of deep learning to fit complicated interactions, especially nonlinear ones, with minimal human engineering, will be of enormous utility in capitalizing on the explosion of data and knowledge. We think there is much room for growth and improvement in disease diagnostics when deep learning is applied. It's important to note that the computational cost of the deep learning-based inter-ensemble approach is higher. To some extent, we could get around this restriction by implementing a feature selection strategy during the data preparation stage; this drastically cut down on the processing time required to make a prediction and increased their accuracy. Selection of characteristics has become increasingly important as the volume and diversity of gene expression information explode.

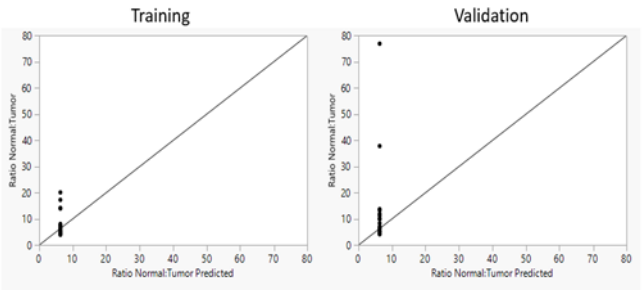

Figure 14. Training and Validation fitting with Ensembled Modeling With likelihood.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 2s
DOI: https://doi.org/10.17762/ijritcc.v10i2s.5918
***Article Received**: 08 October 2022 **Revised**: 07 November 2022 **Accepted**: 26 December 2022*

Generalized Regression for Ratio Normal: Tumor > Normal Maximum Likelihood with Validation Column shown in Fig 14.

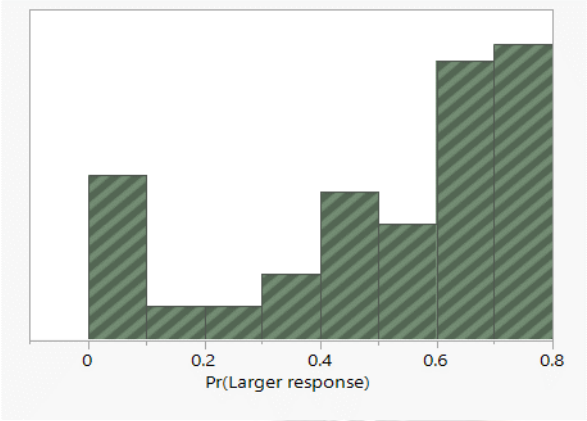

Figure 15. Training and Validation fitting with Ensembled Modeling With likelihood.

Generalized Regression for Ratio Normal: Tumor > Normal Maximum Likelihood with Validation Column > Diagnostic Bundle Shown in Fig15.

IV. CONCLUSION

The global burden of cancer is enormous. Despite the growing popularity of machine learning techniques for cancer forecasting, there is currently no gold standard. In this study, we introduced a multi-model ensemble method to cancer prediction based on deep learning. In particular, we examined data on gene expression levels collected from three different types of tissues: lung, stomach, and breast. We used the Statistical Analysis method to find genes whose expression levels were significantly different between normal and tumor phenotypes, a step that helped us prevent over-fitting in categorization [19]. According to the findings, differentially expressed analysis is essential for lowering the data dimensionality and picking the most relevant pieces of information, which boosts prediction precision while cutting down on computation time. Next, the predictions from many models are fed into a deep neural network that is trained to integrate the inputs into a single, more accurate forecast; this is the multi-model ensemble approach. Combining the verdicts of various classifiers, the majority voting method provides a contrast. The three types of cancer data were examined using the majority voting approach, our suggested multimodal ensemble method, and five individual classifiers. Across several evaluation metrics, the suggested ensemble model is superior to state-of-the-art classifiers and majority voting. Compared to when it is trained independently, the deep learning-based inter-ensemble model has a lower generation error and more data thanks to its use of the predictions from the first stage as features. Additionally, deep learning automatically learns the complex relationships between the classifiers, allowing the ensemble technique to provide more accurate predictions, showing results in Tables 5 and 6.

TABLE 5- Model Summary at Final End

Measure	Training	Validation
Number of rows	43	26
Sum of Frequencies	43	26
-LogLikelihood	122.68094	93.029243
Number of Parameters	1	1
BIC	249.12309	189.31658
AICc	247.45945	188.22515
Generalized RSquare	0	-0.432682

REFERENCES

[1]. L. Zender, A. Villanueva, V. Tovar, D. Sia, D. Y. Chiang, and J. M. Llovet, "Cancer gene discovery in hepatocellular carcinoma," *J. Hepatol.*, vol. 52, no. 6, pp. 921–929, 2010.
[2]. U. Magriples, F. Naftolin, P. E. Schwartz, and M. L. Carcangiu, "High-grade endometrial carcinoma in tamoxifen-treated breast cancer patients.," *J. Clin. Oncol.*, vol. 11, no. 3, pp. 485–490, 1993.
[3]. I. Mármol, C. Sánchez-de-Diego, A. Pradilla Dieste, E. Cerrada, and M. J. Rodriguez Yoldi, "Colorectal carcinoma: a general overview and future perspectives in colorectal cancer," *Int. J. Mol. Sci.*, vol. 18, no. 1, p. 197, 2017.
[4]. S. Chand and others, "A comparative study of breast cancer tumor classification by classical machine learning methods and deep learning method," *Mach. Vis. Appl.*, vol. 31, no. 6, pp. 1–10, 2020.
[5]. C. Natale, G. Z. Leinwand, J. Chiang, J. L. Silberstein, and L. S. Krane, "Reviewing the demographic, prognostic, and treatment factors of primary adenocarcinoma of the bladder: a SEER population-based study," *Clin. Genitourin. Cancer*, vol. 17, no. 5, pp. 380–388, 2019.
[6]. J. Breyer *et al.*, "ESR1, ERBB2, and Ki67 mRNA expression predicts stage and grade of non-muscle-invasive bladder carcinoma (NMIBC)," *Virchows Arch.*, vol. 469, no. 5, pp. 547–552, 2016.
[7]. S. Kuba *et al.*, "Total thyroidectomy versus thyroid lobectomy for papillary thyroid cancer: comparative analysis after propensity score matching: a multicenter study," *Int. J. Surg.*, vol. 38, pp. 143–148, 2017.
[8]. C. Rosty and M. Goggins, "Early detection of pancreatic carcinoma," *Hematol. Clin.*, vol. 16, no. 1, pp. 37–52, 2002.
[9]. R. Suarez-Ibarrola, S. Hein, G. Reis, C. Gratzke, and A. Miernik, "Current and future applications of the machine and deep learning in urology: a review of the literature on urolithiasis, renal cell carcinoma, and bladder and prostate cancer," *World J. Urol.*, vol. 38, no. 10, pp. 2329–2347, 2020.
[10]. K. V. R. Kumar and S. Elias, "Smart Neck-Band for

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 2s
DOI: https://doi.org/10.17762/ijritcc.v10i2s.5918
***Article Received**: 08 October 2022 **Revised**: 07 November 2022 **Accepted**: 26 December 2022*

Rehabilitation of Musculoskeletal Disorders," 2020.

[11]. K. V. R. Kumar, B. R. Devi, M. Sudhakara, G. Keerthi, and K. R. Madhavi, "AI-Based Mental Fatigue Recognition and Responsive Recommendation System," in *Intelligent Computing and Applications*, Springer, 2023, pp. 303–314.

[12]. K. V. R. Kumar and S. Elias, "Real-Time Tracking of Human Neck Postures and Movements," in *Healthcare*, 2021, vol. 9, no. 12, p. 1755.

[13]. G. S. Tandel *et al.*, "A review on a deep learning perspective in brain cancer classification," *Cancers (Basel).*, vol. 11, no. 1, p. 111, 2019.

[14]. M. Olempska, P. A. Eisenach, O. Ammerpohl, H. Ungefroren, F. Fandrich, and H. Kalthoff, "Detection of tumor stem cell markers in pancreatic carcinoma cell lines," *Hepatobiliary Pancreat Dis Int*, vol. 6, no. 1, pp. 92–97, 2007.

[15]. L. Zhou, J. Liu, and F. Luo, "Serum tumor markers for detection of hepatocellular carcinoma," *World J. Gastroenterol. WJG*, vol. 12, no. 8, p. 1175, 2006.

[16]. R. Tabares-Soto, S. Orozco-Arias, V. Romero-Cano, V. S. Bucheli, J. L. Rodr\'\iguez-Sotelo, and C. F. Jiménez-Varón, "A comparative study of machine learning and deep learning algorithms to classify cancer types based on microarray gene expression data," *PeerJ Comput. Sci.*, vol. 6, p. e270, 2020.

[17]. S. D. Bhogaraju, K. V. R. Kumar, P. Anjaiah, J. H. Shaik, and others, "Advanced Predictive Analytics for Control of Industrial Automation Process," in *Innovations in the Industrial Internet of Things (IIoT) and Smart Factory*, I.G.I. Global, 2021, pp. 33–49.

[18]. E. M. K. Reddy, A. Gurrala, V. B. Hasitha, and K. V. R. Kumar, "Introduction to Naive Bayes and a Review on Its Subtypes with Applications," *Bayesian Reason. Gaussian Process. Mach. Learn. Appl.*, pp. 1–14, 2022.

[19]. Ramana, K., Kumar, M. R., Sreenivasulu, K., Gadekallu, T. R., Bhatia, S., Agarwal, P., & Idrees, S. M. (2022). Early prediction of lung cancers using deep saliency capsule and pre-trained deep learning frameworks. *Frontiers in oncology*, *12*.

[20]. Rudra Kumar, M., Pathak, R., & Gunjan, V. K. (2022). Diagnosis and Medicine Prediction for COVID-19 Using Machine Learning Approach. In *Computational Intelligence in Machine Learning* (pp. 123-133). Springer, Singapore.

[21]. Reddy, K. U. K., Shabbiha, S., & Kumar, M. R. (2020). Design of high security smart health care monitoring system using IoT. *Int. J*, *8*.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 2s
DOI: https://doi.org/10.17762/ijritcc.v10i2s.5918
***Article Received**: 08 October 2022 **Revised**: 07 November 2022 **Accepted**: 26 December 2022*

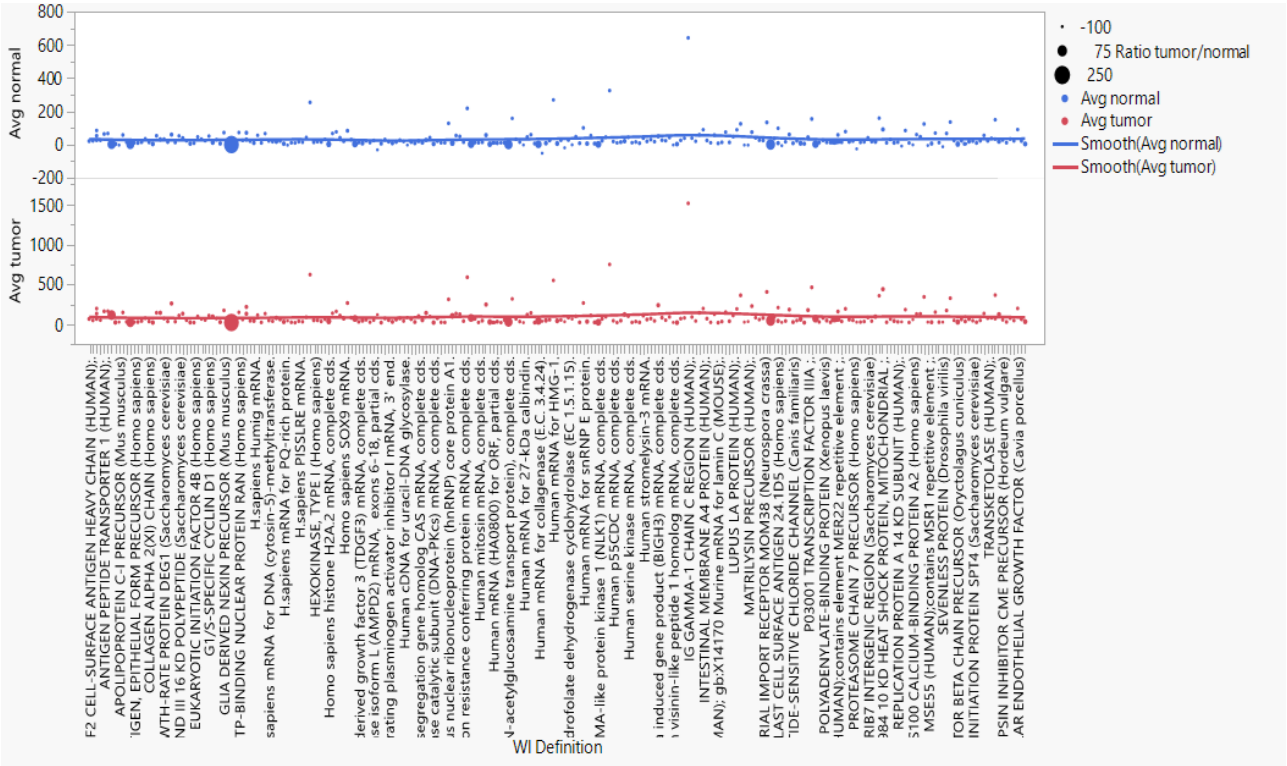

Figure 4: Tumor Vs. Normal cancer conditions.

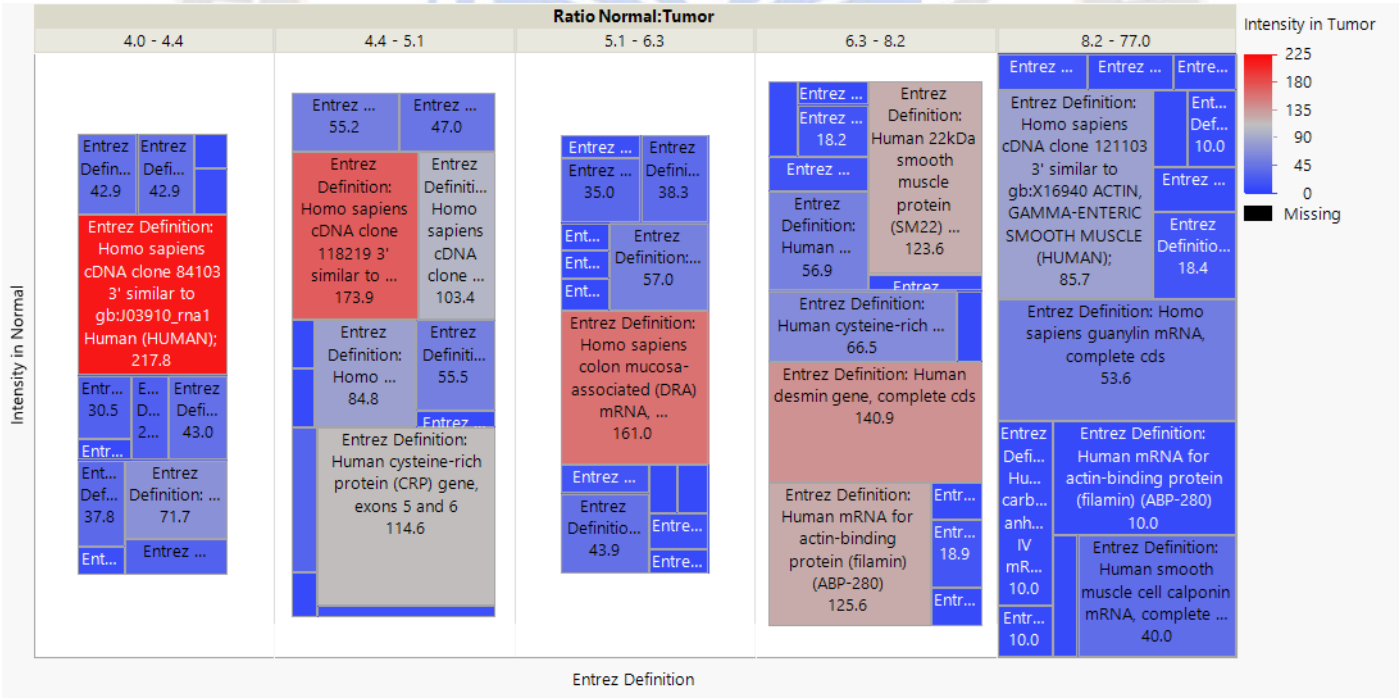

Figure 13: Tumor Vs. Normal cancer conditions. Concerning Ratios

TABLE 6- Final Prediction of Cancer from Tumor conditions.

Term	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%
Intercept	1.8530452	0.1524986	147.65239	<.0001*	1.5541535	2.1519369