Diverse Activations in Ensembles: Odds Ratio Evaluation

Thesis to be submitted in partial fulfillment of the requirements for the degree

of

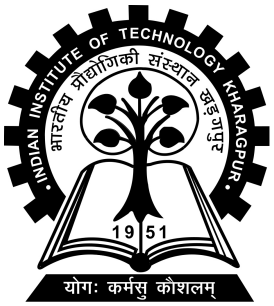
Integrated M.Sc.

by

Karthik Reddy Yeredla
19MA20058

Under the guidance of

Dr.Buddhananda Banerjee



MATHEMATICS

INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

DECLARATION

I certify that

(a) The work contained in this report has been done by me under the guidance of my supervisor.

(b) The work has not been submitted to any other Institute for any degree or diploma.

(c) I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.

(d) Whenever I have used materials (data, theoretical analysis, figures, and text) from other sources, I have given due credit to them by citing them in the text of the thesis and giving their details in the references. Further, I have taken permission from the copyright owners of the sources, whenever necessary.

Date:
Place: Kharagpur

Karthik Reddy Yeredla
19MA20058

Department of Mathematics

Indian Institute of Technology, Kharagpur

India - 721302

CERTIFICATE

This is to certify that we have examined the thesis entitled **Diverse Activations in Ensembles: Odds Ratio Evaluation**, submitted by **Karthik Reddy Yeredla**(Roll Number: *19MA20058*) an undergraduate student of **Department of Mathematics** in partial fulfillment for the award of the degree of Integrated M.Sc.. We hereby accord our approval of it as a study carried out and presented in a manner required for its acceptance in partial fulfillment for the Under Graduate Degree for which it has been submitted. The thesis has fulfilled all the requirements as per the regulations of the Institute and has reached the standard needed for submission.

Dr.Buddhananda Banerjee
Department of Mathematics
Indian Institute of Technology,
Kharagpur

Buddhananda Banerjee
Assistant Professor
Department of Mathematics
IIT Kharagpur, Paschim Medinipur
West Bengal, India. PIN: 721302
Ph : +913222304760

Place: Kharagpur
Date:

ACKNOWLEDGEMENTS

I take this opportunity to express my deepest sense of gratitude and sincere thanks to my MTP supervisor Dr.Buddhananda Banerjee, Department of Mathematics, IIT Kharagpur, for introducing me to this topic and for his invaluable guidance, comments, and suggestions throughout the course of the project. He constantly motivated me to work harder on the problem and provided all the necessary support.

I would also like to thank all my friends and colleagues on the team under Dr.Buddhananda Banerjee for helping me and contributing to the fulfillment of project requirements. I would also like to thank all others whose direct or indirect help has benefited me in doing this project.

Last but not least, I thank my parents and all my family members for the support and motivation they have provided me throughout my life.

Karthik Reddy Yeredla
Department of Mathematics
IIT Kharagpur

Contents

List of Figures

1 Introduction

In the dynamic landscape of machine learning, the continuous quest for model refinement and performance elevation stands as a persistent focal point. This thesis embarks on a captivating exploration into the realm of ensemble learning, specifically focusing on the Bagging technique, a powerful methodology known for its ability to amalgamate diverse models into a unified, robust predictor. However, our attention extends beyond the conventional boundaries of ensemble learning, delving into the nuances of activation layers—a pivotal element often overshadowed in the ensemble paradigm. This research is poised to uncover the intricate interplay between activation functions and ensemble synergy, unraveling hidden patterns that could propel predictive capabilities to unprecedented heights.

As we navigate through this scientific inquiry, our compass is guided by a distinctive evaluative lens - the Odds Ratio. Traditionally utilized in epidemiology and statistics, the Odds Ratio offers a fresh perspective on model assessment. By applying this statistical measure to the ensemble paradigm, we aim to illuminate the comparative strengths and weaknesses of different activation layers within the Bagging framework.

Two distinct datasets serve as the crucible for our experimentation in this research. The first dataset is drawn from the renowned Fashion MNIST [6] collection, comprising 60,000 training images and 10,000 testing images. Each image, represented in grayscale, depicts fashion items categorized into ten distinct classes. The images themselves are square, with a shape of 28x28 pixels, constituting a formidable multiclass classification challenge. Complementing the complexity of Fashion MNIST, the second dataset is a bespoke creation. This binary classification dataset is endowed with 1000 instances, each characterized by 10 features. Notably, this dataset is intentionally structured to exhibit semi-linear separability, introducing a layer of sophistication to our exploration of activation layers within the Bagging ensemble framework. Fashion MNIST Dataset - `www.tensorflow.org/datasets/catalog/fashion_mnist`

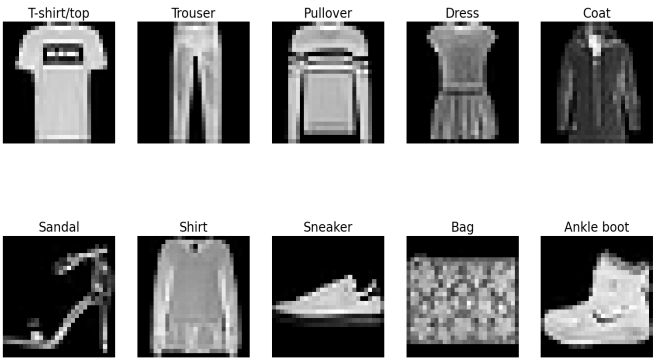


Figure 1: Fashion-MNIST dataset

The significance of this research lies not only in the pursuit of optimal predictive performance but also in expanding the conceptual boundaries of ensemble learning. In doing so, we aim to contribute to the growing body of knowledge that propels the field of machine learning forward.

2 Background

Odds Ratio Analysis, a method widely employed in statistical analysis, serves as a pivotal tool in our exploration. It assesses the likelihood of an event occurring in one group compared to another, providing valuable insights into the relationships within our ensemble learning models. In our experimentation, we leveraged the Mantel-Haenszel Estimator to evaluate the joint performance of models in an ensemble. It goes beyond conventional accuracy metrics by evaluating the combination of individual model predictions, offering a more comprehensive understanding of ensemble dynamics.

2.1 Bootstrap Sampling

Bootstrap sampling[3] is a resampling technique commonly used in statistics to estimate the sampling distribution of a statistic or parameter or to assess the uncertainty of sample estimates. It is particularly valuable in situations where the underlying population distribution is unknown or complex. The process can be formally described as follows:

Let $X = \{x_1, x_2, \ldots, x_N\}$ be an observed dataset with N data points. Bootstrap sampling consists of the following steps:

1. Randomly draw N data points with replacement from the original dataset X, resulting in a resampled dataset X_1^*.

2. Calculate the statistic or parameter of interest, denoted as θ_1^*, based on the resampled dataset X_1^*.

3. Repeat the resampling process a large number of times, typically B iterations, to generate B resampled datasets $\{X_1^*, X_2^*, \ldots, X_B^*\}$ and their corresponding estimates $\{\theta_1^*, \theta_2^*, \ldots, \theta_B^*\}$.

4. Analyze the distribution of the estimated statistics $\{\theta_1^*, \theta_2^*, \ldots, \theta_B^*\}$ to make statistical inferences, construct confidence intervals, or assess the variability of the parameter of interest.

Bootstrap sampling is a fundamental tool in statistics and provides a robust method for making inferences when traditional parametric assumptions may not hold or when analytical solutions are impractical.

2.2 Bagging

A learning set of $\mathcal{D}$ consists of data $\{(y_i, \mathbf{x}_i), i = 1, \ldots, N\}$ where the y's are either class labels or a numerical response. Assume we have a procedure for using this learning set to form a predictor $\hat{f}(\mathbf{x}, \mathcal{D})$ — if the input is $\mathbf{x}$ we predict y by $\hat{f}(\mathbf{x}, \mathcal{D})$. Now, suppose we are given a sequence of learning sets $\{\mathcal{D}_k\}$ each consisting of N independent observations from the same underlying distribution as $\mathcal{D}$. Our mission is to use the $\{\mathcal{D}_k\}$ to get a better predictor than the

single learning set predictor $\hat{f}(\mathbf{x}, \mathcal{D})$. The restriction is that all we are allowed to work with is the sequence of predictors $\{\hat{f}(\mathbf{x}, \mathcal{D}_k)\}$.

If y is numerical, an obvious procedure is to replace $\hat{f}(\mathbf{x}, \mathcal{D})$ by the average of $\hat{f}(\mathbf{x}, \mathcal{D}_k)$ over k, i.e. by $\hat{f}_A(\mathbf{x}) = \mathbb{E}_{\mathcal{D}}[\hat{f}(\mathbf{x}, \mathcal{D})]$ where $\mathbb{E}_{\mathcal{D}}$ denotes the expectation over $\mathcal{D}$, and the subscript A in $\hat{f}_A$ denotes aggregation. If $\hat{f}(\mathbf{x}, \mathcal{D})$ predicts a class $j \in \{1, \ldots, J\}$, then one method of aggregating the $\hat{f}(\mathbf{x}, \mathcal{D}_k)$ is by voting. Let $N_{j,k} = \mathbb{I}\{\hat{f}(\mathbf{x}, \mathcal{D}_k) = j\}$ and take $\hat{f}_A(\mathbf{x}) = \text{argmax}_j \sum_k N_{j,k}$, i.e., the j for which $\sum_k N_{j,k}$ is maximum.

Usually, though, we have a single learning set $\mathcal{D}$ without the luxury of replicates of $\mathcal{D}$. Still, an imitation of the process leading to $\hat{f}_A$ can be done. Take repeated bootstrap samples $\{\mathcal{D}^{(B)}\}$ from $\mathcal{D}$, and form $\{\hat{f}(\mathbf{x}, \mathcal{D}^{(B)})\}$. If y is numerical, take $\hat{f}_B$ as $\hat{f}_B(\mathbf{x}) = av_B \hat{f}(\mathbf{x}, \mathcal{D}^{(B)})$.

If y is a class label, let the $\{\hat{f}(\mathbf{x}, \mathcal{D}^{(B)})\}$ vote to form $\hat{f}_A(\mathbf{x})$. We call this procedure "**b**ootstrap **agg**regat**ing**" and use the acronym **bagging**.[2].

2.3 Odds Ratio

The *OddsRatio* is the ratio of the odds of an event occurring in one group to the odds of it occurring in another group. If the probabilities of the event in each of the groups are p_1 (first group) and p_2 (second group), then the odds ratio is:

$$\frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{p_1/q_1}{p_2/q_2} = \frac{p_1 q_2}{p_2 q_1}, where\ q_i = 1 - p_i,\ i = 1, 2.$$

An odds ratio of 1 indicates that the condition or event under study is equally likely to occur in both groups. An odds ratio greater than 1 indicates that the condition or event is more likely to occur in the first group. An odds ratio of less than 1 indicates that the condition or event is less likely to occur in the first group.

The odds ratio can also be defined in terms of the joint probability distribution of two binary random variables. The joint distribution of binary random variables X and Y can be written

	$Y=1$	$Y=0$
$X=1$	p_{11}	p_{10}
$X=0$	p_{01}	p_{00}

Table 1: Joint Probabilities

	$Y=1$	$Y=0$
$X=1$	$\frac{p_{11}}{p_{11}+p_{10}}$	$\frac{p_{10}}{p_{11}+p_{10}}$
$X=0$	$\frac{p_{01}}{p_{01}+p_{00}}$	$\frac{p_{00}}{p_{01}+p_{00}}$

Table 2: Conditional Probabilities

where p_{11}, p_{10}, p_{01}, and p_{00} are non-negative "cell probabilities" that sum to one. The odds for Y within the two sub-populations defined by X = 1 and X = 0 are defined in terms of the conditional probabilities given X, i.e., $P(Y \mid X)$. Thus the odds ratio is:

$$\frac{p_{11}/(p_{11}+p_{10})}{p_{10}/(p_{11}+p_{10})} \Big/ \frac{p_{01}/(p_{01}+p_{00})}{p_{00}/(p_{01}+p_{00})} = \frac{p_{11}p_{00}}{p_{10}p_{01}}$$

2.4 Mantel-Haenszel Estimator for Common Odds Ratio

Let there be k predictors or weak learners in the ensemble Bagging model for binary classification. Each predictor is represented by $\mathcal{M}_i$ where i = 1,2...,k. The contingency table predictor $\mathcal{M}_i$ is represented as follows:

	$Predicted = 1$	$Predicted = 0$
$Actual = 1$	X_i	Y_i
$Actual = 0$	$N_{i1} - X_i$	$N_{i0} - Y_i$

$$Odds\, ratio\, for\, \mathcal{M}_i = \frac{X_i * (N_{i0} - Y_i)}{Y_i * (N_{i1} - X_i)} \quad and \;\; N_i = N_{i0} + N_{i1}$$

The Mantel-Haenszel estimator[1] is utilized to determine the combined odds ratio for an ensemble of k weak learners. It aggregates individual odds ratios, accounting for potential confounding factors, thereby estimating the overall effect of the entire ensemble model. It is estimated as follows:

$$\psi_{\text{MH}} = \frac{\sum_{i=1}^{k} X_i * (N_{i0} - Y_i)/N_i}{\sum_{i=1}^{k} Y_i * (N_{i1} - X_i)/N_i}$$

2.5 Activation functions

Activation functions play a pivotal role in artificial neural networks, serving as non-linear transformations that introduce complexity and enable the modeling of intricate relationships within data. The choice of activation functions profoundly impacts the network's convergence, generalization, and overall performance, making it a critical consideration in designing and optimizing deep learning models. Common activation functions include the sigmoid, hyperbolic tangent (tanh), and rectified linear unit (ReLU). Here we use the Cumulative Distribution Function (CDF) of Normal Distribution and Cauchy Distribution as activation functions.

Sigmoid Function:

$$\sigma(x) \;=\; \frac{1}{1 + e^{-x}}$$

CDF of Normal Distribution:

$$\Phi(x) = \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{x - \mu}{\sigma\sqrt{2}}\right)\right)$$

$$where \;\; erf\, is\, Gauss\, error\, function, \;\; \mu = Mean \;\; and \;\; \sigma^2 = Variance$$

CDF of Cauchy Distribution:

$$F(x; x_0, \gamma) = \frac{1}{\pi}\arctan\left(\frac{x - x_0}{\gamma}\right) + \frac{1}{2}$$

$$where \;\; x_0 = Mode\, (or)\, Median \;\; and \;\; \gamma = Mean\, Absolute\, Deviation$$

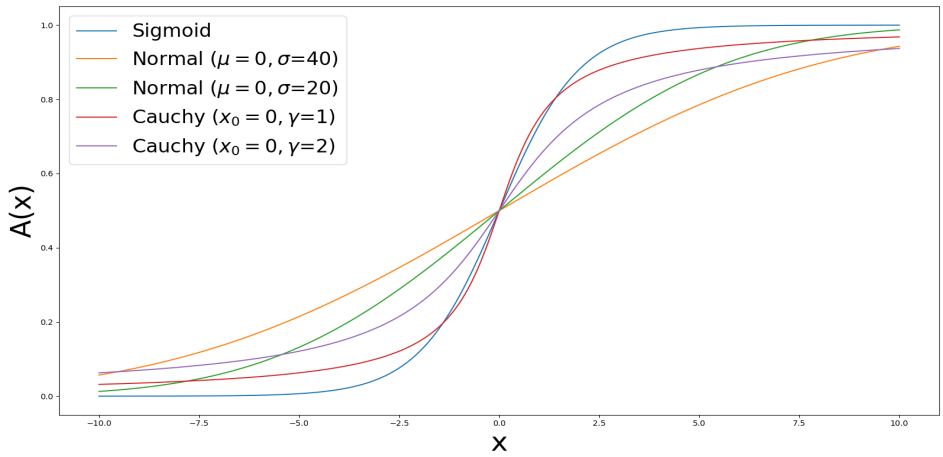


Figure 2: Activation functions

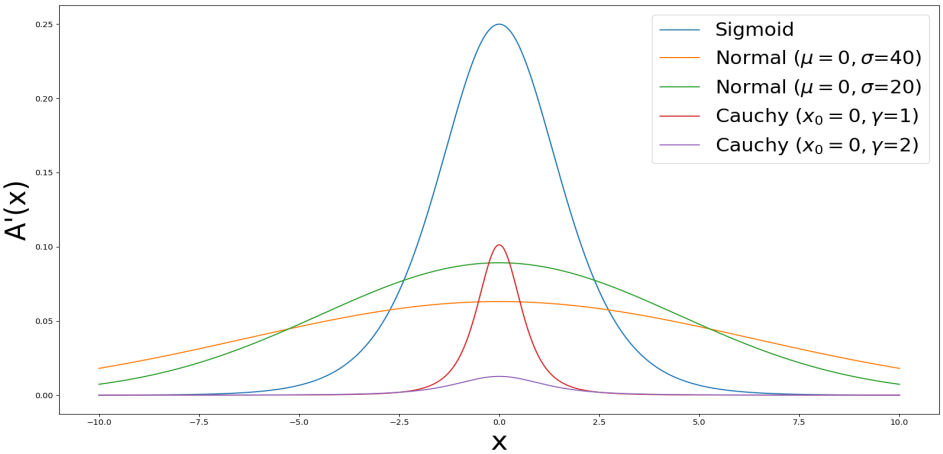


Figure 3: Derivative of Activation functions

2.6 Empirical distribution function

In statistics, an empirical distribution function (commonly also called an empirical cumulative distribution function, ECDF) is the distribution function associated with the empirical measure of a sample. The empirical cumulative distribution function (ECDF) is a non-parametric estimator of the cumulative distribution function (CDF) of a random variable. It is often used when working with sample data to estimate the distribution of the underlying population.

Mathematical Formulation :

Given a dataset $X = \{x_1, x_2, \ldots, x_n\}$, where x_i are the observed data points, the ECDF $F_n(x)$ is defined as follows:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(x_i \leq x)$$

Here, $I(\cdot)$ is the indicator function, which equals 1 if the condition inside the parentheses is true and 0 otherwise. The ECDF is a step function that increases by $1/n$ at each observed

data point.

Practical Implementation :

To compute the ECDF for a given dataset, the following steps are typically taken:

1. **Sort the Data :** Arrange the observed data values in ascending order.
2. **Assign Probabilities :** For each data point, calculate the proportion of data points less than or equal to it.
3. **Plot the ECDF :** Plot the sorted data points against their corresponding empirical cumulative probabilities.

2.7 Two-Sample Kolmogorov-Smirnov test

This test [4] is employed to compare two independent samples and determine if they are drawn from the same distribution.

Mathematical Formulation :

Given two independent samples $X = \{x_1, x_2, \ldots, x_m\}$ and $Y = \{y_1, y_2, \ldots, y_n\}$, the two-sample KS statistic D is defined as:

$$D = \sup_x |F_m(x) - F_n(x)|$$

Here, $F_m(x)$ and $F_n(x)$ are the empirical cumulative distribution functions (ECDFs) of the first and second samples, respectively. The critical value(p-value) of the KS statistic can be compared to tabulated values or obtained through simulation methods.

Practical Implementation :

Conducting a two-sample KS test involves the following steps:

1. **Compute the ECDFs :** Calculate the empirical cumulative distribution functions for both samples.
2. **Calculate the KS Statistic :** Determine the maximum vertical distance between the two sample ECDFs.
3. **Compare with Critical Values :** Assess the significance level of the KS statistic by comparing it with critical values.

3 Experiment Set-up

3.1 Dataset

In the pursuit of binary classification within the Fashion MNIST dataset, two visually akin categories were chosen - Category 2: Pullover and Category 6: Shirt. This decision was rooted in the desire to create a discerning challenge for our models by selecting classes that share subtle visual resemblances. By focusing on the Pullover and Shirt categories, we aim to investigate the efficacy of activation layers in distinguishing between these closely related apparel types. The inherent similarities between the chosen categories introduce a layer of complexity, pushing the boundaries of our binary classification task.

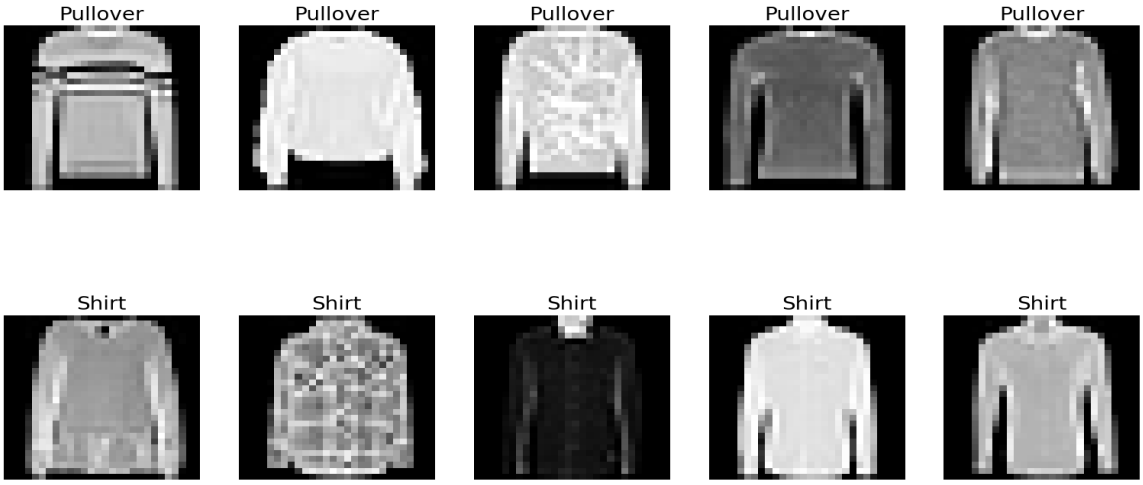


Figure 4: Pullover(Category 2) and Shirt(Category 6) in Fashion MNIST dataset

3.2 Weak Learner and its Architecture

In the Bagging ensemble framework, the weak learner serves as the foundational building block, designed to make predictions with limited accuracy. Common examples of weak learners include decision stumps (shallow decision trees with only one split), linear models with minimal complexity, or simple rule-based classifiers.

In the context of this research, a neural network is chosen as the weak learner owing to its flexibility and capacity to capture intricate patterns. The selected neural network architecture takes an input dimensionality of 784, corresponding to flattened 28x28 images from the Fashion MNIST dataset. The neural network comprises a singular hidden layer containing 64 neurons. This layer is equipped with activation functions, influencing the model's capacity to capture complex relationships. The output layer, featuring a single neuron, encapsulates the binary classification nature of our task. Two types of experiments are being done, depending on which type of experiment is being done the activation layers in the output neuron are changed.

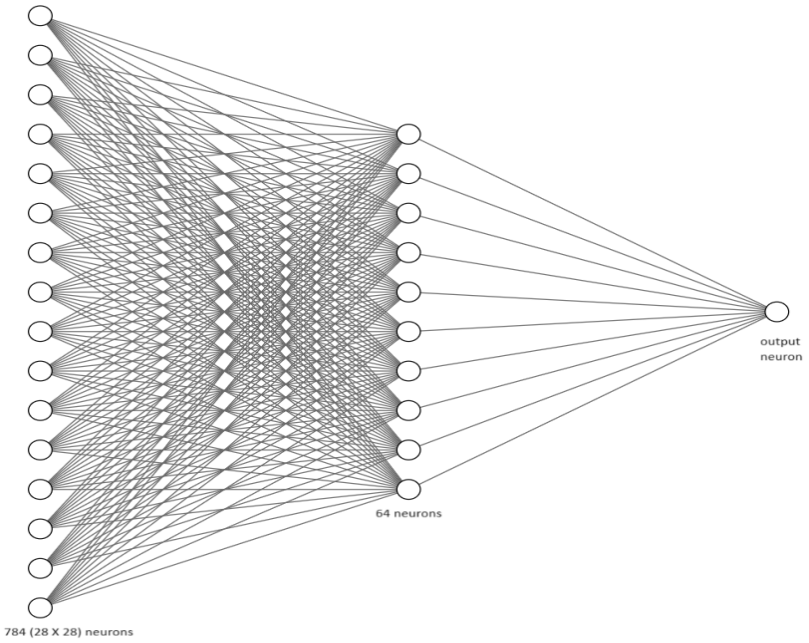


Figure 5: Architecture of neural network used in Bagging Ensemble

3.3 Ensemble Model's Architecture

The ensemble model comprises 10 neural network-based weak learners, unified by a standardized architecture. The crux of our experimentation lies in the activation functions at the final output neuron. Across our varied experiments, this output neuron showcases diverse activation functions, including sigmoid, cumulative distribution function (CDF) of the Normal Distribution, and the CDF of the Cauchy Distribution. This deliberate divergence in activation functions serves as the focal point of our investigation, enabling a distinct exploration of the impact of distinct activation layers within the Bagging ensemble framework. Two types of experiments are as follows:

Type I : All 10 weak learners have the activation function in the output neuron as a Sigmoid.

Type II : Three Different activation functions are used among the 10 weak learners.

$Activation\,Function(In\,output\,layer)$	$Number\,of\,Weak\,Learners$
$Sigmoid$	2
$CDF\,of\,Normal\,Distribution\,(\mu = 0\,, \sigma = 20)$	2
$CDF\,of\,Normal\,Distribution\,(\mu = 0\,, \sigma = 40)$	2
$CDF\,of\,Cauchy\,Distribution\,(x_0 = 0\,, \gamma = 1)$	2
$CDF\,of\,Cauchy\,Distribution\,(x_0 = 0\,, \gamma = 2)$	2

Table 3: Distribution of Activation Layers Among Weak Learners in the Ensemble Model

Now the whole ensemble model is trained on the Fashion MNIST train dataset and evaluated on the Fashion MNIST test dataset. The process is repeated 1000 times to get a sample distribution of the evaluation metric(Odd's ratio) on the train and test dataset. The results are shown in the following section.

4 Results

We selected two visually similar categories, specifically Category 2: Pullover and Category 6: Shirt from the Fashion MNIST dataset, for binary classification using an ensemble model. The ensemble model was iteratively executed 1000 times to observe the distribution of common odds ratio evaluation metrics. Here are the findings based on the test data.

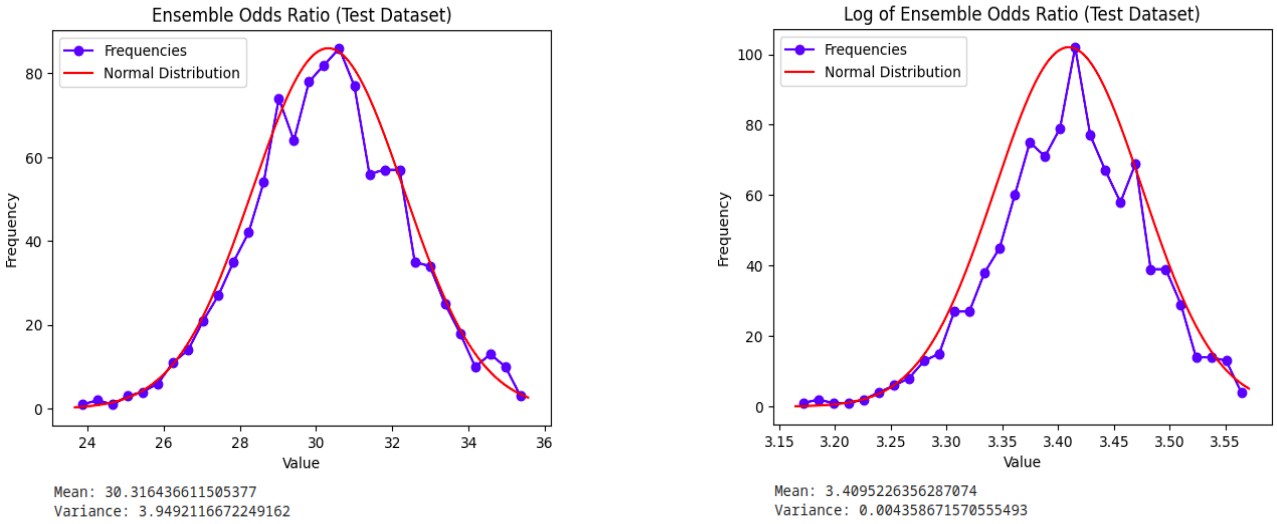


Figure 6: Odds ratio distribution on Test Dataset(Type - I)

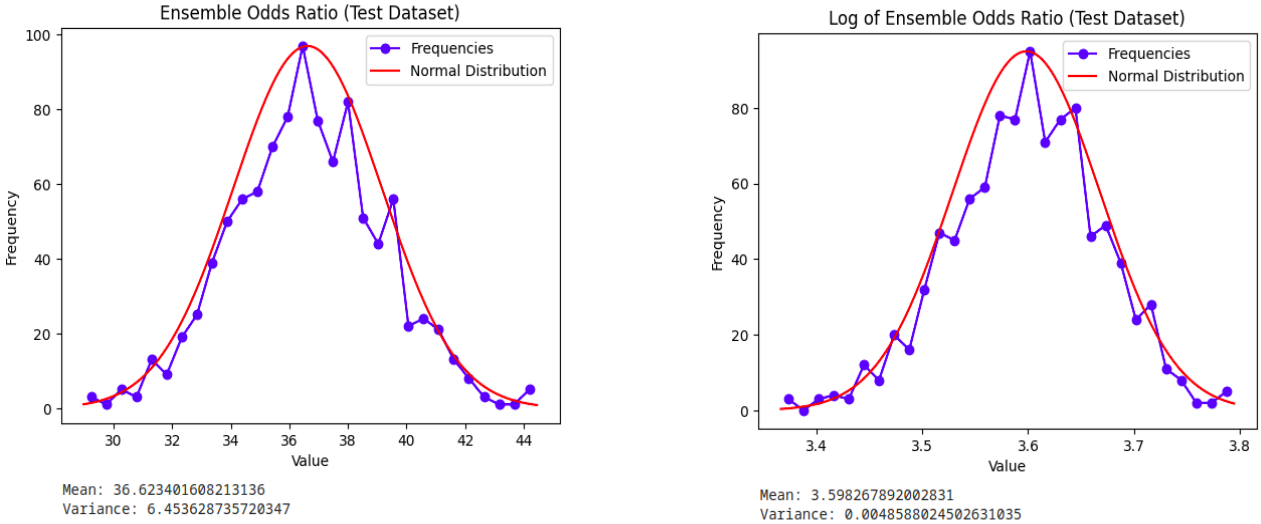


Figure 7: Odds ratio distribution on Test Dataset(Type - II)

The mean value of the odds ratio on the test dataset increased from 30.31 to 36.62 when we transitioned from using all sigmoid activation functions to incorporating different activation functions. This indicates an improvement in the ensemble model's performance with the introduction of diverse activation functions. Notably, along with the increase in mean, there was also an observed rise in variance.

The KS statistic value of 0.63 and an extremely low p-value of $\mathbf{2.36 \times 10^{-188}}$, obtained from a 2-sample KS test, indicate a significant difference between the distributions of odds

ratio values under the two scenarios (all sigmoid activation functions vs. different activation functions). This statistical test strengthens the evidence that the change in activation functions has a substantial impact on the ensemble model's behavior.

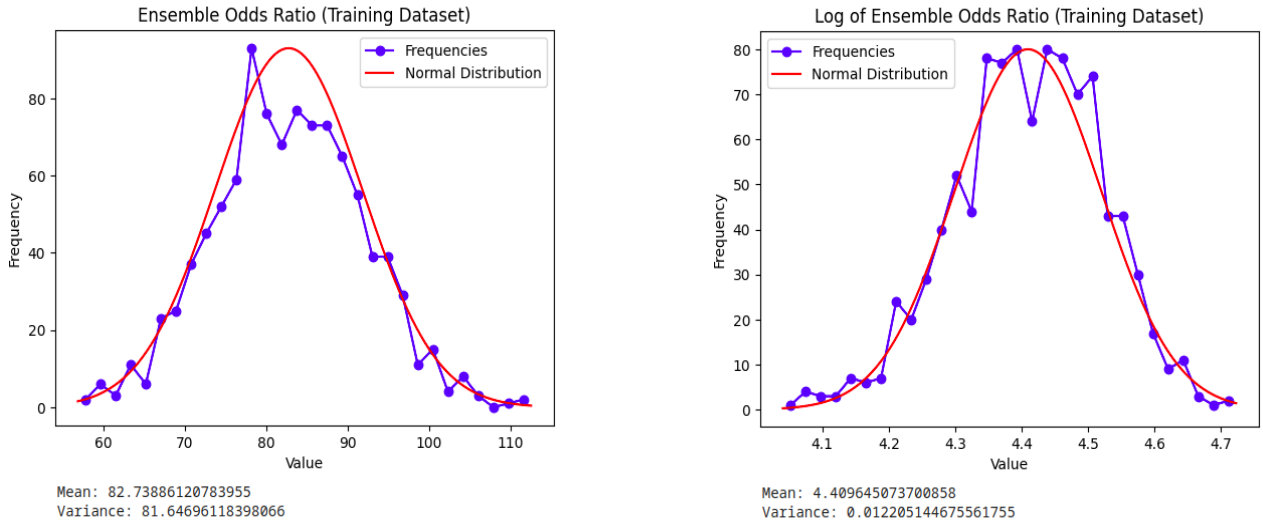


Figure 8: Odds ratio distribution on Train Dataset(Type - I)

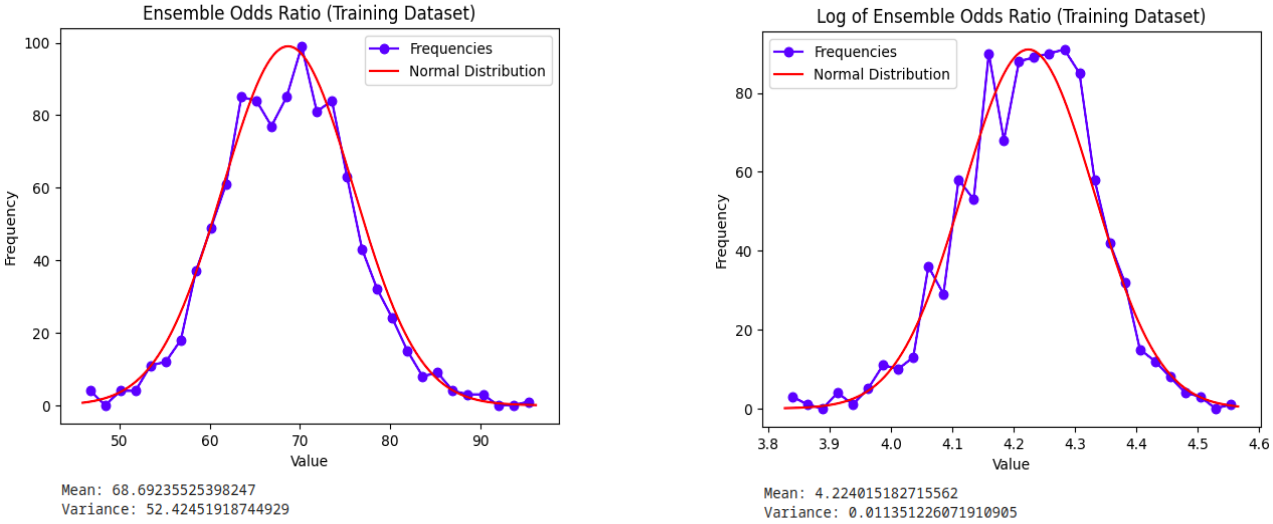


Figure 9: Odds ratio distribution on Train Dataset(Type - II)

The decrease in the mean value of the training set, from 82.73 to 68.69, when changing activation functions from all sigmoid to different activation functions. Simultaneously the test set performance metrics show an increase, which implies that the model is not only fitting the training data well but also generalizing effectively to new and unseen instances. We can also observe that the distributions are similar to Normal Distribution.

5 Further Works

In the scope of future developments, this study aims to pioneer the creation of a novel non-parametric data-driven activation function for neural networks. By harnessing the inherent characteristics and patterns within the data itself, the proposed activation function seeks to dynamically adjust its behavior without relying on fixed parameters. The envisioned activation function aspires to redefine the conventional paradigms by allowing neural networks to autonomously shape their activation responses based on the intrinsic data distributions encountered during training.

The research conducted by [5]Inturrisi et al. highlights the utility of piecewise linear activation functions in augmenting the performance of deep neural networks. These adaptive piecewise linear units dynamically adjust their behavior based on learned parameters, enabling them to capture intricate data patterns efficiently. By citing this work, the report acknowledges the significance of parametric data-driven activation functions in enhancing the adaptability and learning capabilities of neural networks, providing a solid foundation for future endeavors in this domain.

References

[1] Buddhananda Banerjee and Atanu Biswas. On closeness of the mantel–haenszel estimator and the profile likelihood-based estimator of the common odds ratio from multiple 2×2 tables. *Statistics and Probability Letters*, 82(11):1990–1993, 2012.

[2] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, August 1996.

[3] B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, January 1979.

[4] Joseph L. Hodges. The significance probability of the smirnov two-sample test. *Arkiv för Matematik*, 3:469–486, 1958.

[5] Jordan Inturrisi, Sui Yang Khoo, Abbas Kouzani, and Riccardo Pagliarella. Piecewise linear units improve deep neural networks. *arXiv preprint arXiv:2108.00700*, 08 2021.

[6] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747, 2017.