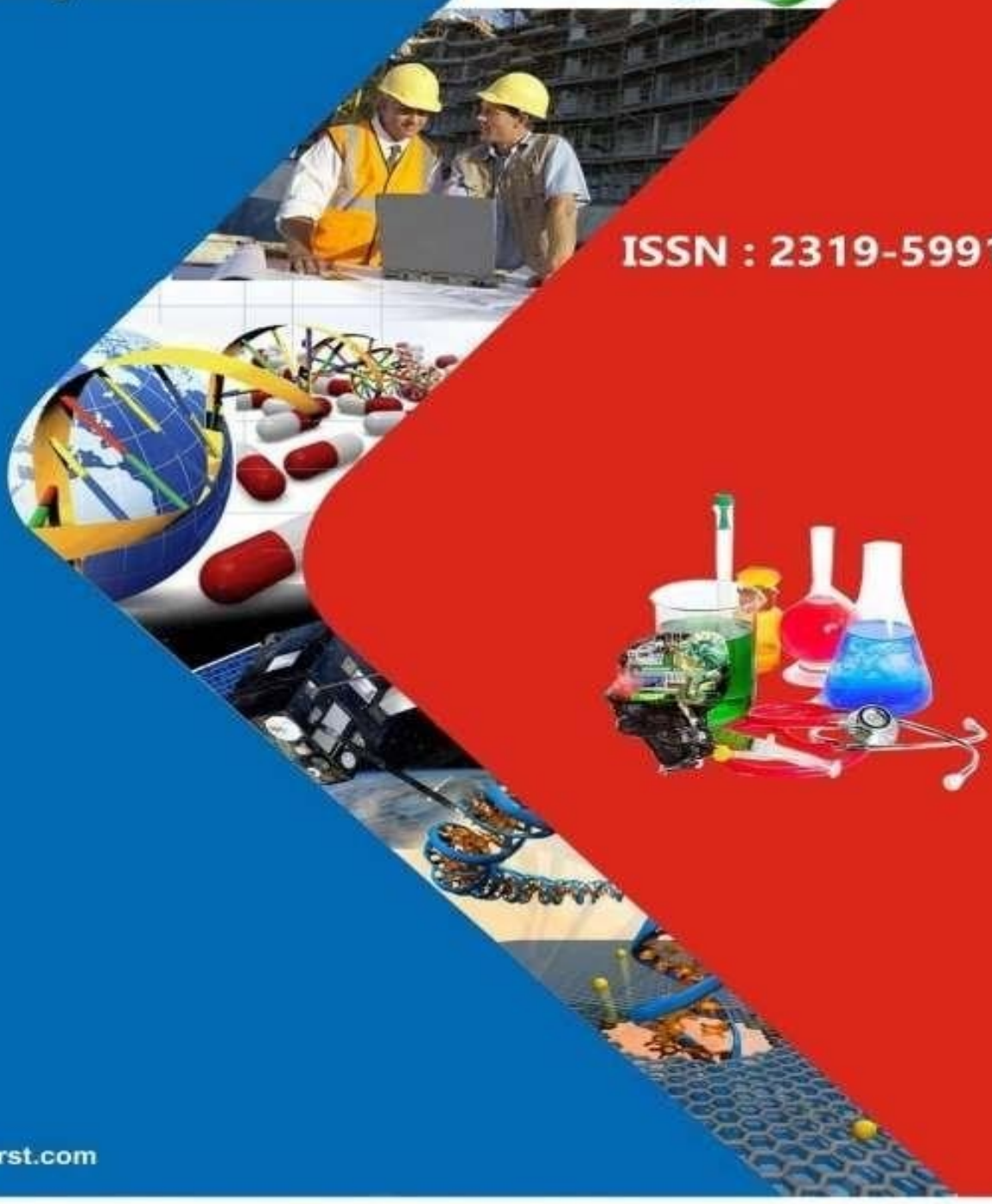
International Journal of Engineering Research and Science & Technology

IJERST

ISSN : 2319-5991



www.ijerst.com

Email: editor@ijerst.com or editor.ijerst@gmail.com

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

https://doi.org/10.62643/ijerst.2025.v21.i2.pp903-912

Vol. 21, Issue 2, 2025

SMART DEFENSE: A STUDENT-TEACHER FRAMEWORK FOR ADAPTIVE ADVERSARIAL ROBUSTNESS

[1]K. Thrivikram, MCA Student, Department of MCA

[2] Dr.K.Pavan Kumar, Ph.D, Associate Professor, Department of MCA

[12]Dr KV Subba Reddy Institute of Technology, Dupadu, Kurnool

ABSTRACT

Deep neural networks' (DNNs') dependability and security depend on defence against hostile assaults. The most advanced defence techniques available today are very resilient to hostile assaults. Nevertheless, these defence strategies are unable to differentiate between normal instances (NEs) and adversary examples (AEs). As a consequence, they execute categorisation using the same defence procedure for both instances, which degrades NE performance. In this research, we offer a new defence approach based on the student-teacher framework that can identify AEs and apply the defence process exclusively to AEs, hence minimising the deterioration of classification performance for NEs. We train the student network to predict the undistorted hidden layer features of the teacher network (target DNN), focussing on the fact that adversarial assaults would always succeed if the hidden layer features are distorted. As a consequence, our approach can identify AEs by comparing the hidden layer features of the teacher and student networks, and then use the penultimate layer characteristics that the student network predicted to retrieve the classification result of AEs. By conducting comprehensive tests on relevant image classification benchmark datasets, including as CIFAR-10, CIFAR-100, and TinyImagenet, we show that our approach outperforms state-of-the-art techniques in both detection and defence. Additionally, we demonstrate that our approach produces strong detection and defence results for a completely white-box assault, which presumes an attacker is aware of every detail of our detection and defence system.

I. INTRODUCTION

In a number of tasks, such as image classification [1], [2], [3], object recognition [4], semantic segmentation [5], and natural language processing [6], deep neural networks (DNNs) have shown exceptional performance. Nevertheless, DNNs are susceptible to adversarial assaults that alter their inputs in subtle ways [7], [8]. Specifically, repeatedly optimising perturbations using adversarial assaults like Projected Gradient Descent (PGD) [9], Carlini and Wagner (CW) [10], Deep Fool (DF) [11], and Auto Attack (AA) [12] may almost completely eliminate the DNN classifier's accuracy in image classification tasks. The weaknesses of DNNs are further highlighted by the numerous adversarial attack techniques that have been put forth for tasks involving text classification, object detection, face recognition, point cloud classification, and more, in addition to image classification [13], [14], [15], and [16]. Applying DNNs to safety-critical fields like autonomous driving is seriously threatened by this [17].

There are two categories of techniques for dealing with misclassification brought on by adversarial attacks: detection and defence. Detection is a reactive technique that finds adversarial instances (AEs) to filter incorrectly categorised outcomes. In contrast to normal instances (NEs), the hidden layer characteristics for Aes show an abnormal distribution, according to a number of studies [18], [19], [20], [21], [22], [23], and [24]. In light of this, detection techniques identify abnormal traits that arise in AEs and use them to carry out detection [25], [26], [27], and [28]. These techniques can

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
https://doi.org/10.62643/ijerst.2025.v21.i2.pp903-912
Vol. 21, Issue 2, 2025

consistently stop adversarial attack-induced misclassification without lowering NE classification performance because they use detection models apart from classification models. The drawback of detection techniques is that they are unable to get accurate AE categorisationfindings.

Conversely, defence strategies concentrate on correctly categorising AEs. By creating AEs for every batch and including them into the training process, adversarial training—the most representative defence strategy—increases the resilience of DNNs against adversarial assaults [9], [29]. Numerous adversarial training techniques have been put forward, and the approach that makes use of a significant number of extra datasets has lately produced state-of-the-art results [30], [31], [32], and [33]. Adversarial purification is another strategy for competitive defence. To guarantee that the model generates accurate classification results, these techniques purify adversarial perturbations in the input using generative models like GANs [34], [35], energy-based models (EBMs) [36], and diffusion models [37].

The defence approach's ability to accurately classify AEs makes it effective. These techniques, however, are unable to identify whether the input is an AE. As a consequence, they are unable to determine if an adversarial attack is now a danger, and because both NEs and AEs are subject to the same defence procedure, NE classification performance degrades. Robustness for AEs and classification accuracy for NEs are traded off in the context of adversarial training [29]. Because it purifies NEs even in the absence of adversarial perturbations, adversarial purification also lowers classification accuracy for NEs. A defence must also be able to detect if an adversarial assault is now a danger in real-world situations. Moreover, the classification performance reduction for NEs is deadly for practicality, since AEs are often less common than NEs.

We provide a framework-based approach for detecting and defending AEs in order to tackle this issue. By first identifying whether the input is AE or not, and then only using defence mechanisms on inputs that are found to be AE in order to recover the proper classification result, our approach may minimise the deterioration of classification performance for NEs. Our approach uses the teacher network as the classifier and the student network as the model that has been trained to predict the teacher network's undistorted hidden layer attributes. Consequently, the degree of distortion in the hidden layer features is represented by the variations between the teacher and student networks' hidden layer features. We may use this difference as a detection score to identify AEs since adversarial assaults always cause distortion of hidden layer features, which taints DNN predictions. Additionally, we may get the classification results for the identified AEs by utilising the penultimate layer characteristics that the student network predicted. The final layer characteristics of the student and teacher networks for NEs and AEs on the CIFAR-10 dataset are visualised in Figure 1. Figure 1's top row demonstrates how the penultimate layer characteristics of the teacher network for NEs and the student network for AEs are located in close proximity to one another. To increase accuracy against the AEs, we also develop a preprocessing method called restoration assault. A disturbance caused by a restoration assault causes the gap between the student and instructor networks' hidden layer characteristics to widen. Next, we combine this disturbance with the input that was found to be AE. The restoration attack forces the student network to forecast hidden layer features that may provide accurate classification results when the teacher network generates distorted hidden layer features for AEs. When AEs are preprocessed by restoration attack, the distribution for the penultimate layer

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
https://doi.org/10.62643/ijerst.2025.v21.i2.pp903-912
Vol. 21, Issue 2, 2025

characteristics of the student network is shown in the bottom row of Figure 1. These student network's penultimate layer features are more closely grouped inside the appropriate class area and show more distance from the instructor network's distorted features.

The following are the study's contributions:

• To the best of our knowledge, we provide the first approach that combines defence and detection. Therefore, by applying a defence strategy just to identified AEs, our technique may address the decrease of classification performance for NEs.

• To further improve the classification performance for AEs, we develop a restoration attack, an input preprocessing strategy.

• Compared to current algorithms, our approach delivers state-of-the-art performance in both detection and defence tasks in trials on relevant image benchmark datasets: CIFAR-10, CIFAR-100, and TinyImageNet.

• By attaining excellent performance for a completely white-box assault, in which the attacker has complete knowledge of our detection and defence mechanism as well as the classifier, we show the resilience of our approach.

II. LITERATURE SURVEY

"Classification of ImageNet using deep convolutional neural networks,"

G. E. Hinton, I. Sutskever, and A. Krizhevsky,

To categorise the 1.2 million high-resolution photos in the ImageNet LSVRC-2010 competition into the 1000 distinct classes, we trained a large, deep convolutional neural network. We significantly outperformed the prior state-of-the-art with top-1 and top-5 error rates of 37.5% and 17.0% on the test data, respectively. With 60 million parameters and 650,000 neurones, the neural network is made up of three fully-connected layers, a final 1000-way softmax, and five convolutional layers, some of which are followed by max-pooling layers. We used non-saturating neurones and a very effective GPU version of the convolution process to speed up training. We used a newly created regularisation technique called "dropout" to lessen overfitting in the fully-connected layers, and it worked really well. In the ILSVRC-2012 competition, we also submitted a variation of this model and won with a top-5 test error rate of 15.3%, whereas the second-best entry had a top-5 error rate of 26.2%.

"Very deep convolutional networks to recognise images on a large scale,"

In this study, K. Simonyan and A. Zisserman examine how the depth of a convolutional network affects its accuracy in the context of large-scale image recognition. Our primary contribution is a detailed analysis of networks with increasing depth utilising an architecture with extremely tiny (3x3) convolution filters, demonstrating that raising the depth to 16-19 weight layers may significantly enhance the prior-art configurations. These results served as the foundation for our team's entry to the 2014 ImageNet Challenge, where we placed first in the localisation track and second in the classification track. Additionally, we demonstrate how successfully our representations generalise to additional datasets, achieving state-of-the-art outcomes. To encourage further study on the use of deep visual representations in computer vision, we have made our two top-performing ConvNet models publicly accessible.

"Image recognition using deep residual learning,"

J. Sun, S. Ren, X. Zhang, and K. He,

Training deeper neural networks is more challenging. In order to facilitate the training of networks that are far deeper than those previously used, we provide a residual learning approach. Rather than learning unreferenced functions, we explicitly reformulate the layers as learning residual functions with reference to the layer inputs. We provide thorough empirical proof that these residual networks are simpler to optimise and that a significant increase in depth may improve accuracy. We test residual nets with a depth of up to 152 layers on the ImageNet dataset, which is 8× deeper than VGG nets [40] but still less complicated. On the ImageNet test set, an ensemble of these residual nets achieves an error

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
https://doi.org/10.62643/ijerst.2025.v21.i2.pp903-912
Vol. 21, Issue 2, 2025

of 3.57%. On the ILSVRC 2015 classification task, this result took first place. Analysis of CIFAR-10 with 100 and 1000 layers is also presented. For many visual identification tasks, the depth of representations is crucial. We achieve a 28% relative improvement on the COCO object identification dataset, just because of our very deep representations. Our contributions to the ILSVRC & COCO 2015 competitions1 were built on deep residual nets, and we took first place in the ImageNet detection, ImageNet localisation, COCO detection, and COCO segmentation tasks.

"Faster R-CNN: Towards region proposal networks for real-time object detection,"

J. Sun, R. Girshick, K. He, and S. Ren,

To make educated guesses about item locations, state-of-the-art object detection networks rely on region proposal algorithms. The running duration of these detection networks has been shortened by developments like SPPnet [1] and Fast R-CNN [2], revealing region proposal computing as a bottleneck. In this study, we provide a Region Proposal Network (RPN) that allows for almost cost-free region proposals by sharing full-image convolutional features with the detection network. A fully convolutional network that predicts object limits and objectness scores at every point at the same time is called an RPN. Fast R-CNN uses the high-quality region suggestions produced by the RPN's end-to-end training for detection. Using the increasingly trendy notion of neural networks with 'attention' processes, we further combine RPN and Fast R-CNN into a single network by sharing their convolutional features. The RPN component instructs the united network where to search. With just 300 recommendations per picture, our detection method achieves state-of-the-art object recognition accuracy on PASCAL VOC 2007, 2012, and MS COCO datasets, while operating at a frame rate of 5 fps (including all stages) on a GPU for the extremely deep VGG-16 model [3]. Faster R-CNN and RPN are the cornerstones of the first-place winning submissions in a number of tracks in the ILSVRC and COCO 2015 competitions. The code is now openly accessible.

For semantic segmentation, fully convolutional networks are used.

T. Darrell, J. Long, and E. Shelhamer,

Convolutional networks are strong visual models that produce feature hierarchies. We demonstrate that convolutional networks outperform the state-of-the-art in semantic segmentation when trained end-to-end, pixel-to-pixel. Our main breakthrough is the creation of "fully convolutional" networks, which efficiently infer and learn from inputs of any size and generate outputs of the same size. We provide an explanation of how fully convolutional networks are used to spatially dense prediction problems, define and describe the space of these networks, and make links to previous models. Modern classification networks, such as GoogLeNet [32], the VGG net [31], and AlexNet [20], are modified into fully convolutional networks, and their learnt representations are applied to the segmentation task by fine-tuning [3]. Next, we establish a skip architecture that generates precise and comprehensive segmentations by fusing appearance data from a shallow, fine layer with semantic data from a deep, coarse layer. With an inference time of less than 0.5 seconds for a typical picture, our fully convolutional network delivers state-of-the-art segmentation of PASCAL VOC (20% relative improvement to 62.2% mean IU on 2012), NYUDv2, and SIFT Flow.BERT: Deep bidirectional transformer pre-training for language comprehension,

M.-W. Chang, K. Lee, K. Toutanova, J. Devlin,

Bidirectional Encoder Representations from Transformers, or BERT for short, is a novel language representation paradigm that we offer. BERT is intended to pre-train deep bidirectional representations from unlabelled text by jointly conditioning on both left and right context in all layers, in contrast to current language representation models. Therefore, without requiring significant task-specific architectural changes, the pre-trained BERT model may be

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
https://doi.org/10.62643/ijerst.2025.v21.i2.pp903-912
Vol. 21, Issue 2, 2025

optimised with only one more output layer to provide state-of-the-art models for a variety of tasks, including question answering and language inference.

BERT is both strong experimentally and theoretically. On eleven natural language processing tasks, it achieves new state-of-the-art results, such as increasing the GLUE score to 80.5% (7.7% point absolute improvement), MultiNLI accuracy to 86.7% (4.6% absolute improvement), SQuAD v1.1 question answering Test F1 towards 93.2 (1.5 point absolute improvement), and SQuAD v2.0 Test F1 towards 83.1 (5.1 point absolute improvement).

III. SYSTEM ANALYSIS & DESIGN

EXISTING SYSTEM

The Mahalanobis Distance (MD) is used in another research [27] to identify adverse events. In order to identify AEs, they employ the MD as input variables of the detection model after calculating the MD of each hidden layer under the assumption that hidden layer features follow the Gaussian distribution of each class. According to Reference [28], AEs are more sensitive than NEs to changes in the decision boundary in areas with significant curvature. They use this realisation to train a dual classifier that uses a weighted average wavelet transform to utilise the sensitivity discrepancy information in the input pictures. After that, they use the encoding vectors of the dual and primal classifiers as input features to train a detection model.

A representative defence technique called adversarial training creates and trains adversarial events (AEs) for every batch in order to acquire robust classifiers [9]. Reference [30] examines a variety of variables that might boost adversarial training's efficacy and demonstrates how adversarial training performs noticeably better with a large number of extra datasets. Additionally, [31], [38] uses a generative model for data augmentation to attain state-of-the-art performance. As described by [39], a number of approaches have also been put forward to lower the computational cost of adversarial training; nonetheless, performance and efficiency are still not equal. Another effective defensive technique is adversarial purification, which use generative models to eliminate adversarial assaults by eliminating input perturbations. Numerous techniques have been put forward that make use of various generative models, such as GANs [34], autoregressive models [40], and EBMs [36], [41]. DiffPure, which use diffusion models as the purification model, was recently presented [38].

Disadvantages

- Using this architecture, the system was unable to identify and resist AEs since it lacked a consistent mechanism.
- The student network was not trained to predict the undistorted hidden layer properties of the teacher network, which is the classifier, using an established approach.

PROPOSED SYSTEM

We provide a technique that can identify and defend AEs, based on a student-teacher paradigm. By first identifying whether the input is AE or not, and then only using defence mechanisms on inputs that are found to be AE in order to recover the proper classification result, our approach may minimise the deterioration of classification performance for NEs. Our approach uses the teacher network as the classifier and the student network as the model that has been trained to predict the teacher network's undistorted hidden layer attributes. Consequently, the degree of distortion in the hidden layer features is represented by the variations between the teacher and student networks' hidden layer features. We may use this difference as a detection score to identify AEs since adversarial assaults always cause distortion of hidden layer features, which taints DNN predictions. Additionally, we may retrieve the classification results for the identified AEs by utilising the penultimate layer characteristics that the student network predicted.

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

https://doi.org/10.62643/ijerst.2025.v21.i2.pp903-912

Vol. 21, Issue 2, 2025

To increase accuracy against AEs, we develop a preprocessing method termed restoration assault. A disturbance caused by a restoration assault causes the gap between the student and instructor networks' hidden layer characteristics to widen. Next, we combine this disturbance with the input that was found to be AE. The restoration attack forces the student network to forecast hidden layer features that may provide accurate classification results when the teacher network generates distorted hidden layer features for AEs.

Advantages

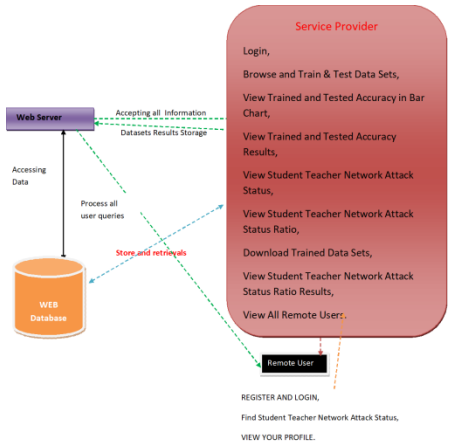
• To the best of our knowledge, we provide the first approach that combines defence and detection. Therefore, by implementing a defence procedure only for identified AEs, our approach may mitigate classification performance decrease for NEs.

• To further improve the classification performance for AEs, we develop a restoration attack, an input preprocessing strategy.

• Compared to current algorithms, our approach delivers state-of-the-art performance in both detection and defence tasks in trials on relevant image benchmark datasets: CIFAR-10, CIFAR-100, and TinyImageNet.

• By attaining excellent performance for a completely white-box assault, in which the attacker has complete knowledge of our detection and defence mechanism as well as the classifier, we show the resilience of our approach.

SYSTEM ARCHITECTURE



IV. IMPLEMENTATIONS

Modules

Service Provider

The Service Provider must use a working user name and password to log in to this module. He can do a number of tasks after successfully logging in, including browsing and training and testing data sets. See the Bar Chart for Trained and Tested Accuracy. View Accuracy Tested and Trained Results, Download Trained Data Sets, View Student Teacher Network Attack Status, and View Student Teacher Network Attack Status Ratio View the results of the Student Teacher Network Attack Status Ratio. See Every Remote User.

View and Authorize Users

The administrator may see a list of all registered users in this module. Here, the administrator may see the user's information, like name, email, and address, and they can also grant the user permissions.

Remote User

A total of n users are present in this module. Before beginning any actions, the user needs register. Following registration, the user's information will be entered into the database. Following a successful registration, he must use his password and authorised user name to log in. Following a successful login, the user will do

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
https://doi.org/10.62643/ijerst.2025.v21.i2.pp903-912
Vol. 21, Issue 2, 2025

tasks such as registering and logging in, determining the status of student teacher network attacks, Examine your profile.

ALGORITHMS

Gradient boosting

One machine learning method for classification and regression problems is gradient boosting. Usually decision trees, it provides a prediction model in the form of an ensemble of weak prediction models.[1] [2] The resultant technique, known as gradient-boosted trees, often performs better than random forest when a decision tree is the weak learner.Like other boosting techniques, a gradient-boosted trees model is constructed step-by-step; however, it goes one step further by allowing optimisation of an arbitrary differentiable loss function.

K-Nearest Neighbors (KNN)

- Using a similarity measure, this simple but very effective classification system is non-parametric, lazy learning, and doesn't "learn" until the test sample is provided.
- We use the training data to determine the K-nearest neighbours of any fresh data that has to be classified.

Example

- Learning based on instances, which also operates sluggishly because instances near the input vector for test or prediction may take time to occur in the training dataset;
- Training dataset comprises k-closest examples in feature space;
- Feature space means, space with categorisation variables (non-metric variables).

Logistic regression Classifiers

The relationship between a collection of independent (explanatory) factors and a categorical dependent variable is examined using logistic regression analysis. When the dependent variable simply has two values, like 0 and 1 or Yes and No, the term logistic regression is used. When the dependent variable contains three or more distinct values, such as married, single, divorced, or widowed, the technique is sometimes referred to as multinomial logistic regression. While the dependent variable's data type differs from multiple regression's, the procedure's practical application is comparable.

When it comes to categorical-response variable analysis, logistic regression and discriminant analysis are competitors. Compared to discriminant analysis, many statisticians believe that logistic regression is more flexible and appropriate for modelling the majority of scenarios. This is due to the fact that, unlike discriminant analysis, logistic regression does not presume that the independent variables are regularly distributed.

Both binary and multinomial logistic regression are calculated by this software for both category and numerical independent variables. Along with the regression equation, it provides information on likelihood, deviance, odds ratios, confidence limits, and quality of fit. It does a thorough residual analysis that includes diagnostic residual plots and reports. In order to find the optimal regression model with the fewest independent variables, it might conduct an independent variable subset selection search. It offers ROC curves and confidence intervals on expected values to assist in identifying the optimal classification cutoff point. By automatically identifying rows that are not utilised throughout the study, it enables you to confirm your findings.

Naïve Bayes

The supervised learning technique known as the "naive bayes approach" is predicated on the straightforward premise that the existence or lack of a certain class characteristic has no bearing on the existence or nonexistence of any other feature.

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

https://doi.org/10.62643/ijerst.2025.v21.i2.pp903-912

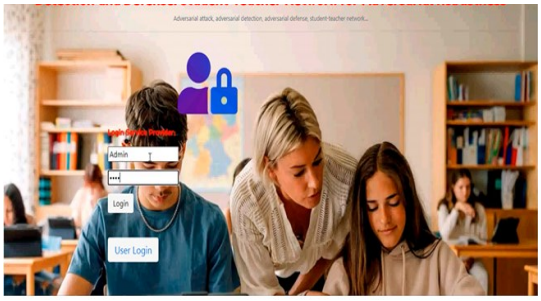
Vol. 21, Issue 2, 2025

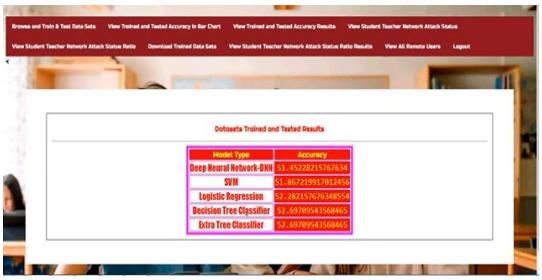
However, it seems sturdy and effective in spite of this. It performs similarly to other methods of guided learning. Numerous explanations have been put forward in the literature. We emphasise a representation bias-based explanation in this lesson. Along with logistic regression, linear discriminant analysis, and linear SVM (support vector machine), the naive bayes classifier is a linear classifier. The technique used to estimate the classifier's parameters (the learning bias) makes a difference.

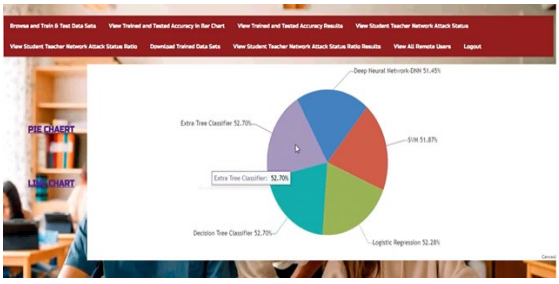
Although the Naive Bayes classifier is commonly used in research, practitioners who want to get findings that are useful do not utilise it as often. On the one hand, the researchers discovered that it is very simple to build and apply, that estimating its parameters is simple, that learning occurs quickly even on extremely big datasets, and that, when compared to other methods, its accuracy is rather excellent. The end users, however, do not comprehend the value of such a strategy and do not get a model that is simple to read and implement.

As a consequence, we display the learning process's outcomes in a fresh way. Both the deployment and comprehension of the classifier are simplified. We discuss several theoretical facets of the naive bayes classifier in the first section of this lesson. Next, we use Tanagra to apply the method on a dataset. We contrast the outcomes (the model's parameters) with those from other linear techniques including logistic regression, linear discriminant analysis, and linear support vector machines. We see that the outcomes are quite reliable. This helps to explain why the strategy performs well when compared to others. We employ a variety of tools (Weka 3.6.0, R 2.9.2, Knime 2.1.1, Orange 2.0b, and RapidMiner 4.6.0) on the same dataset in the second section. Above all, we make an effort to comprehend the outcomes.

V. SCREEN SHOTS





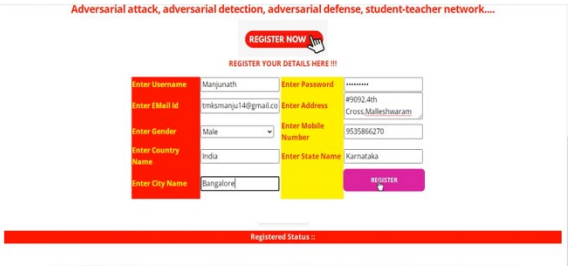


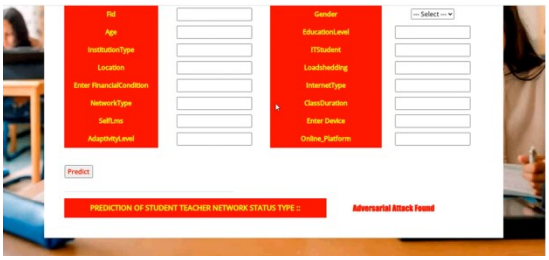
International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

https://doi.org/10.62643/ijerst.2025.v21.i2.pp903-912

Vol. 21, Issue 2, 2025





VI. CONCLUSION

In this study, we use a student-teacher context to present a unique adversarial robustness technique that combines defence and detection. Our approach uses a student network trained to predict undistorted hidden layer features to estimate the amount of distortion in the hidden layer features, hence detecting AEs. Next, we use the penultimate layer properties that the student network predicted to reclassify the discovered cases. Our approach provides state-of-the-art performance in both detection and defence, as shown by extensive trials on typical image classification benchmark datasets. Specifically, we enhance defence performance and attain strong defence performance even against completely white-box assaults by using the preprocessing approach we develop, restoration attack.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, ‘‘ImageNet classification with deep convolutional neural networks,’’ in Proc. 26th Annu. Conf. Neural Inf. Process. Syst., Lake Tahoe, NV, USA, 2012,pp. 1106–1114.

[2] K. Simonyan and A. Zisserman, ‘‘Very deep convolutional networks for large-scale image recognition,’’ in Proc. 3rd Int. Conf. Learn. Represent.,San Diego, CA, USA, 2015.

[3] K. He, X. Zhang, S. Ren, and J. Sun, ‘‘Deep residual learning for image recognition,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, Jun. 2016, pp. 770–778, doi:10.1109/CVPR.2016.90.

[4] S. Ren, K. He, R. Girshick, and J. Sun, ‘‘Faster R-CNN: Towards real-time object detection with region proposal networks,’’ in Proc. Adv. Neural Inf.Process. Syst., vol. 28, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, 2015, pp. 1–9.

[5] J. Long, E. Shelhamer, and T. Darrell, ‘‘Fully convolutional networks for semantic segmentation,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 3431–3440.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, ‘‘BERT: Pre-training of deep bidirectional transformers for language understanding,’’ 2018,arXiv:1810.04805.

[7] I. J. Goodfellow, J. Shlens, and C. Szegedy, ‘‘Explaining and harnessing adversarial examples,’’ in Proc. 3rd Int. Conf. Learn. Represent., San Diego, CA, USA, 2015.

[8] C. Szegedy,W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, ‘‘Intriguing properties of neural networks,’’ in Proc. 2nd Int. Conf. Learn. Represent., Banff, AB, Canada, 2014,pp. 1–10.

[9] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, ‘‘Towards deep learning models resistant to adversarial attacks,’’ in Proc. 6th Int. Conf. Learn. Represent., Vancouver, BC, Canada, Apr. 2018. [Online].Available: https://openreview.net/forum?id=rJzIBfZAb

[10] N. Carlini and D. Wagner, ‘‘Adversarial examples are not easily detected:Bypassing ten

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
https://doi.org/10.62643/ijerst.2025.v21.i2.pp903-912
Vol. 21, Issue 2, 2025

detection methods,'' in Proc. 10th ACM Workshop Artif. Intell. Secur., Dallas, TX, USA, Nov. 2017, pp. 3–14, doi:10.1145/3128572.3140444.
[11] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, ''DeepFool: A simple and accurate method to fool deep neural networks,'' in Proc. IEEE Conf.Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, Jun. 2016, pp. 2574–2582, doi: 10.1109/CVPR.2016.282.
[12] F. Croce and M. Hein, ''Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks,'' in Proc. Int. Conf. Mach. Learn., 2020, pp. 2206–2216. [Online]. Available: http://proceedings.mlr.press/v119/croce20b.html
[13] L. Song, X. Yu, H.-T. Peng, and K. Narasimhan, ''Universal adversarial attacks with natural triggers for text classification,'' 2020,arXiv:2005.00174.
[14] M. Yin, S. Li, C. Song, M. S. Asif, A. K. Roy-Chowdhury, and S. V. Krishnamurthy, ''ADC: Adversarial attacks against object detection that evade context consistency checks,'' in Proc. IEEE/CVF Winter Conf.Appl. Comput. Vis. (WACV), Jan. 2022, pp. 3278–3287.
[15] C.-Y. Lin, F.-J. Chen, H.-F. Ng, and W.-Y. Lin, ''Invisible adversarial attacks on deep learning-based face recognition models,'' IEEE Access,vol. 11, pp. 51567–51577, 2023.
[16] J. Zhang, Y. Dong, J. Zhu, J. Zhu, M. Kuang, and X. Yuan, ''Improving transferability of 3D adversarial attacks with scale and shear transformations,''Inf. Sci., vol. 662, Mar. 2024, Art. no. 120245.
[17] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, ''Robust physical-world attacks on deep learning models,'' 2017, arXiv:1707.08945.
[18] K. Grosse, P. Manoharan, N. Papernot, M. Backes, and P. McDaniel, ''On the (statistical) detection of adversarial examples,'' 2017, arXiv:1702.06280.
[19] F. Carrara, F. Falchi, R. Caldelli, G. Amato, R. Fumarola, and R. Becarelli,''Detecting adversarial example attacks to deep neural networks,'' in Proc.15th Int. Workshop Content-Based Multimedia Indexing, Florence, Italy, Jun. 2017, p. 38, doi: 10.1145/3095713.3095753.
[20] X. Li and F. Li, ''Adversarial examples detection in deep networks with convolutional filter statistics,'' in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Venice, Italy, Oct. 2017, pp. 5775–5783, doi: 10.1109/ICCV.2017.615.
[21] L. Smith and Y. Gal, ''Understanding measures of uncertainty for adversarial example detection,'' in Proc. 34th Conf. Uncertainty Artif.Intell., Monterey, CA, USA, Aug. 2018, pp. 560–569. [Online]. Available: http://auai.org/uai2018/proceedings/papers/207.pdf
[22] J. Liu,W. Zhang, Y. Zhang, D. Hou, Y. Liu, H. Zha, and N. Yu, ''Detection based defense against adversarial examples from the steganalysis point of view,'' in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Long Beach, CA, USA, Jun. 2019, pp. 4820–4829.
[23] P. Sperl, C.-Y. Kao, P. Chen, X. Lei, and K. Böttinger, ''DLA: Dense-layer-analysis for adversarial example detection,'' in Proc. IEEE Eur. Symp. Secur. Privacy, Genoa, Italy, Sep. 2020, pp. 198–215, doi:10.1109/EuroSP48549.2020.00021.
[24] K. Roth, Y. Kilcher, and T. Hofmann, ''The odds are odd: A statistical test for detecting adversarial examples,'' in Proc. 36th Int. Conf. Mach. Learn., 2019, pp. 5498–5507. [Online]. Available: http://proceedings.mlr.press/v97/roth19a.html
[25] R. Feinman, R. R. Curtin, S. Shintre, and A. B. Gardner, ''Detecting adversarial samples from artifacts,'' 2017, arXiv:1703.00410.