# DATA POISONING ATTACKS ON FEDERATED MACHINE LEARNING

**Dr. Waseema Masood** Associate Professor Department of CSE Deccan College of Engineering and Technology Affiliated to Osmania University Hyderabad,Telangana :: waseema@deccancollege.ac.in

**Zeesha** PG Scholar Department of CSE Deccan College of Engineering and Technology Affiliated to Osmania University Hyderabad,Telangana :: zeesha2262@gmail.com

***Abstract*— Federated Deep learning which enables resource constrained node devices (e.g., mobile phones and IoT devices) to learn a shared model while keeping the training data local, can provide privacy, security and economic benefits by designing an effective communication protocol. However, the communication protocol amongst different nodes could be exploited by attackers to launch data poisoning attacks, which has been demonstrated as a big threat to most Deep learning models. In this paper, we attempt to explore the vulnerability of federated Deep learning. More specifically, we focus on attacking a federated multi-task learning framework, which is a federated learning framework via adopting a general multi-task learning framework to handle statistical challenges. We formulate the problem of computing optimal poisoning attacks on federated multi-task learning as a bilevel program that is adaptive to arbitrary choice of target nodes and source attacking nodes. Then we propose a Server Upload Download Upload Download iPhone10:20 AM iPhone10:20 AM Upload Download Clean Data. Clean Data Clean Data Injected Data Node 1 Node 2 Injected Data Node n Fig. 1. The demonstration of our data poisoning attack model on federated novel systems-aware optimization method, ATTack on Federated Learning (AT2FL), which is efficiency to derive the implicit gradients for poisoned data, and further compute optimal attack strategies in the federated Deep learning. Our work is an earlier study that considers issues of data poisoning attack for federated learning. To the end, experimental results on real-world datasets show that federated multi-task learning model is very sensitive to poisoning attacks, when the attackers either directly poison the target nodes or indirectly poison the related nodes by exploiting the communication protocol.**

*Index Terms*— **Federated Machine Learning, Data Poisoning**

## I. INTRODUCTION

Deep learning has been widely-applied into a broad array of applications, e.g., spam filtering and natural gas price prediction. Among these applications, the reliability or security of the Deep learning system has been a great concern, including adversaries. For example, for product recommendation system , researchers can either rely on public crowd-sourcing platform, e.g., Amazon Mechanical Turk or Taobao, or private teams to collect training datasets. However, both of these above methods have the opportunity of being injected corrupted or poisoned data by attackers. To improve the robustness of real-world Deep learning systems, it is critical to study how well Deep learning performs under the poisoning attacks. For the attack strategy on Deep learning methods, it can be divided into two categories: causative attacks and.

In recent past the world has witnessed the largest number of mortalities due to cardiovascular disease (CVD) as compared to other diseases. It is mostly caused due to high blood pressure, abnormal pulse rate, high cholesterol, diabetes and high glucose level. The mental stress, physical inactivity, smoking habits, alcohol addiction, obesity, family history, unhealthy diets and lack of physical activities are the associated risk factors which may lead to the occurrence of the CVD. The CAD is a type of CVD and is mostly diagnosed by the ECG signal which captures the abnormality of the heart. But due to small amplitude of ECG signal, the clinicians often fail to identify its abnormality. Therefore, development of reliable DL based models for early detection and robust classification of CAD is a challenging task exploratory attacks , where exploratory attacks influence learning via controlling over training data, and exploratory attacks can take use of misclassifications without affecting training. However, previous researches on poisoning attacks focus on the scenarios that training samples are collected in a centralized location, or the training samples are sent to a centralized location via a distributed network, e.g., support vector Deeps, autoregressive models and collaborative f iltering. However, none of the current works study poisoning attacks on federated Deep learning, where the training data are distributed across multiple devices (e.g., users' mobile devices: phones/tablets), and may be privacy sensitive. To further improve its robustness, in this paper, our work explores how to attack the federated Deep learning. For the federated Deep learning, its main idea is to build Deep learning models based on data sets that are distributed across multiple devices while preventing data leakage. Although recent improvements have been focusing on overcoming the statistical challenges (i.e., the data collected across the network are in a non-IID manner, where the data on each node are generated by a distinct distribution) or improving privacy preserving, the attempt that makes federated learning more reliability under poisoning attacks, is still scarce. For example, consider several different e-commerce companies in a same region, and the target is to establish a prediction model for product purchase based on user and product information, e.g., user's browsing and purchasing history. The attackers can control a prescribed number of user accounts and inject the poisoning data in a direct manner. Furthermore, due to the communication protocol existing amongst different companies, 2 IEEE Internet of Things Journal, Volume:9, Issue:13,01. July.2022 Furthermore, poisoning attack is investigated on many Deep this protocol also opens a door for the attacker to indirectly affect the inaccessible target nodes, which is also not addressed by existing poisoning methods whose training data are collected in a centralized.

## II. LITERATURE REVIEW

**Scott Alfeld, Xiaojin Zhu, and Paul Barford. Data poisoning attacks against autoregressive models. In Thirtieth AAAI Conference on Artificial Intelligence, 2016..**

Forecasting models play a key role in money-making ventures in many different markets. Such models are often trained on data from various sources, some of which may be untrustworthy. An actor in a given market may be incentivized to drive predictions in a certain direction to their own benefit. Prior analyses of intelligent adversaries in a Deep-learning context have focused on regression and classification. In this paper we address the non-iid setting of time series forecasting. We consider a forecaster, Bob, using a fixed, known model and a recursive forecasting method. An adversary, Alice, aims to pull Bob's forecasts toward her desired target series, and may exercise limited influence on the initial values fed into Bob's model. We consider the class of linear autoregressive models, and a flexible framework of encoding Alice's desires and constraints. We describe a method of calculating Alice's optimal attack that is computationally tractable, and empirically demonstrate its effectiveness compared to random and greedy baselines on synthetic and real-world time series data. We conclude by discussing defensive strategies in the face of Alice-like adversaries.

**Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How to backdoor federated learning. arXiv preprint arXiv:1807.00459, 2018**

We propose an Federated learning enables thousands of participants to construct a deep learning model without sharing their private training data with each other. For example, multiple smartphones can jointly train a next-word predictor for keyboards without revealing what individual users type. We demonstrate that any participant in federated learning can introduce hidden backdoor functionality into the joint global model, e.g., to ensure that an image classifier assigns an attacker-chosen label to images with certain features, or that a word predictor completes certain sentences with an attacker-chosen word. We design and evaluate a new model-poisoning methodology based on model replacement. An attacker selected in a single round of federated learning can cause the global model to immediately reach 100% accuracy on the backdoor task. We evaluate the attack under different assumptions for the standard federated-learning tasks and show that it greatly outperforms data poisoning. Our generic constrain-and-scale technique also evades anomaly detection-based defenses by incorporating the evasion into the attacker's loss function during training..

**Marco Barreno, Blaine Nelson, Anthony D Joseph, and J Doug Tygar. The security of Deep learning. Deep Learning, 81(2):121–148, 2010.**

Deep learning’s ability to rapidly evolve to changing and complex situations has helped it become a fundamental tool for computer security. That adaptability is also a vulnerability: attackers can exploit Deep learning systems. We present a taxonomy identifying and analyzing attacks against Deep learning systems. We show how these classes influence the costs for the attacker and defender, and we give a formal structure defining their interaction. We use our framework to survey and analyze the literature of attacks against Deep learning systems. We also illustrate our taxonomy by showing how it can guide attacks against SpamBayes, a popular statistical spam filter. Finally, we discuss how our taxonomy suggests new lines of defenses.

**Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Return of the devil in the details: Delving deep into convolutional nets. arXiv preprint arXiv:1405.3531, 2014.**

We included 155 adult PLN mutation carriers and 155 age- and sex-matched control subjects. Twenty-one PLN mutation carriers (13.4%) were classified as symptomatic (symptoms of heart failure or malignant ventricular arrhythmias). The data set was split into training and testing sets using 4-fold cross-validation. Multiple models were developed to discriminate between PLN mutation carriers and control subjects. For comparison, expert cardiologists classified the same data set. The best performing models were validated using an external PLN p.Arg14del mutation carrier data set from Murcia, Spain (n = 50). We applied occlusion maps to visualize the most contributing ECG regions.

**Ingrid Daubechies, Massimo Fornasier, and Ignace Loris. Accelerated projected gradient method for linear inverse problems with sparsity constraints. journal of fourier analysis and applications, 14(5-6):764 792, 2008.**

We investigate a family of poisoning attacks against Support Vector Deeps (SVM). Such attacks inject specially crafted training data that increases the SVM's test error. Central to the motivation for these attacks is the fact that most learning algorithms assume that their training data comes from a natural or well-behaved distribution. However, this assumption does not generally hold in security-sensitive settings. As we demonstrate, an intelligent adversary can, to some extent, predict the change of the SVM's decision function due to malicious input and use this ability to construct malicious data. The proposed attack uses a gradient ascent strategy in which the gradient is computed based on properties of the SVM's optimal solution. This method can be kernelized and enables the attack to be constructed in the input space even for non-linear kernels. We experimentally demonstrate that our gradient ascent procedure reliably identifies good local maxima of the non-convex validation error surface, which significantly increases the classifier's test error..

**Ling Huang, Anthony D Joseph, Blaine Nelson, Benjamin IP Rubinstein, and JD Tygar. Adversarial Deep learning. In Proceedings of the 4th ACM workshop on Security and artificial intelligence, pages 43–58. ACM, 2011.**

The latest generation of Convolutional Neural Networks (CNN) have achieved impressive results in challenging benchmarks on image recognition and object detection, significantly raising the interest of the community in these methods. Nevertheless, it is still unclear how different CNN methods compare with each other and with previous state-of-the-art shallow representations such as the Bag-of-Visual-Words and the Improved Fisher Vector. This paper conducts a rigorous evaluation of these new techniques, exploring different deep architectures and comparing them on a common ground, identifying and disclosing important implementation details. We identify several useful properties of CNN-based representations, including the fact that the dimensionality of the CNN output layer can be reduced significantly without having an adverse effect on performance. We also identify aspects of deep and shallow methods that can be successfully shared. In particular, we show that the data augmentation techniques commonly applied to CNN-based methods can also be applied to shallow methods, and result in an analogous performance boost. Source code and models to reproduce the experiments in the paper is made publicly available.

## III. EXISTING METHODS:

For the data poisoning attacks, it has become an urgent research field in the adversarial machine learning, in which the target is against machine learning algorithms. The earlier attempt that investigates the poisoning attacks on support vector machines (SVM), where the adopted attack uses a gradient ascent strategy in which the gradient is obtained based on properties of the SVM’s optimal solution.

## IV. PROPOSED SYSTEM

The system proposes a bilevel optimization framework to compute optimal poisoning attacks on federated machine learning. To our best knowledge, this is an earlier attempt to explore the vulnerability of federated machine learning from the perspective of data poisoning. The proposed system derives an effective optimization method, i.e., Attack on Federated Learning (AT2FL), to solve the optimal attack problem, which can address systems challenges associated with federated machine learning. The proposed system demonstrates the empirical performance of our optimal attack strategy, and our proposed AT2FL algorithm with several real-world datasets. The experiment results indicate that the communication protocol among multiple nodes opens a door for attacker to attack federated machine learning.

## METHODOLOGY:

- **Data COLLECTION:**
  In this dataset we collect two fruit images dataset with set of images of apple and banana with more than 500 images in each folder. For showing data poisoning attack we are using avcado images inside apple folder and train dataset. From these images image pixel values are taken as features and folder name is taken as labels.

- **Pre-processing:**

Pre-processing is a procedure adopted to enhance the quality of images and increase visualization. In fruit imaging, image processing is a crucial phase that helps to improve the images quality. This can be one of the most critical factors in achieving good results and accuracy in next phases of proposed methodology. fruit images may contain a different issue that may lead to poor and low visualization of the image. If the images are poor or of low quality, it may lead to unsatisfactory results. During preprocessing phase, we performed background elimination, elimination of non-essential blood supplies, image enhancement, and noise removal.

- **Train-Test Split and Model FITTING:**

  Now, we divide our dataset into training and testing data. Our objective for doing this split is to assess the performance of our model on unseen data and to determine how well our model has generalized on training data. This is followed by a model fitting which is an essential step in the model building process.

- **Model Evaluation and Predictions:**

  This is the final step, in which we assess how well our model has performed on testing data using certain scoring metrics, I have used 'accuracy score' to evaluate my model. First, we create a model instance, this is followed by fitting the training data on the model using a fit method and then we will use the predict method to make predictions on x_test or the testing data, these predictions will be stored in a variable called y_test_hat. For model evaluation, we will feed the y_test and y_test_hat into the accuracy_score function and store it in a variable called test_accuracy, a variable that will hold the testing accuracy of our model. We followed these steps for a variety of classification algorithm models and obtained corresponding test accuracy score

- **Prediction show Data poison attack:**

  In this method we develop a web application using flask frame work when user uploads apple image federated learning algorithm will predict as apple if input images is banana output will be banana but if avocado image which is in apple folder is given as input it will predict as attack.

**ARCHITECTURE:**

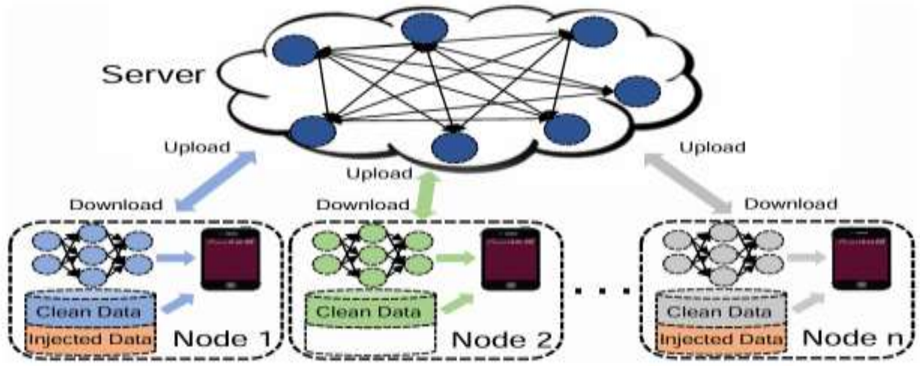


Figure 1. System Architecture

Figure 1 illustrates the framework of the proposed method The demonstration of our data poisoning attack model on federated machine learning, where different colors denote different nodes, and there are n nodes in this federated learning system. Some nodes are injected by corrupted/poisoned data, and some nodes are only with clean data.

**EfficientNet:** A family of convolutional neural networks that uses a compound scaling method to balance network depth, width, and resolution for improved efficiency and performance.

**Inception:** A deep learning architecture that utilizes parallel convolutional layers of different sizes to capture various spatial features in an image, known for its "Inception modules."

**ResNet:** A deep neural network architecture that introduces residual connections or "skip connections" to allow training of very deep networks by mitigating the vanishing gradient problem..

**FLOW DIAGRAM:**

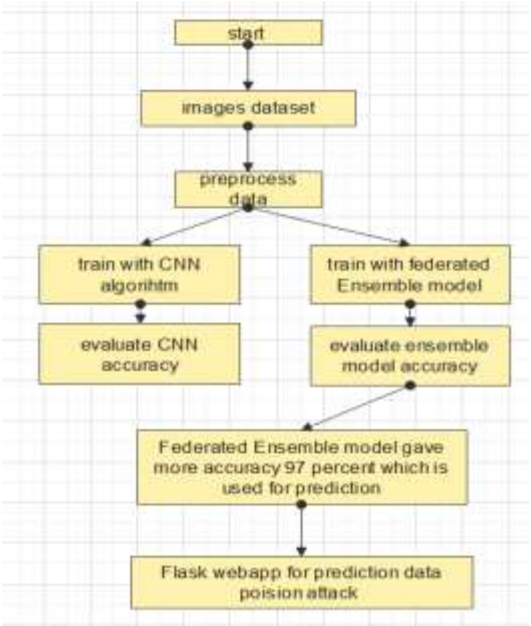

start
images dataset
preprocess data
train with CNN algorihtm
train with federated Ensemble model
evaluate CNN accuracy
evaluate ensemble model accuracy
Federated Ensemble model gave more accuracy 97 percent which is used for prediction
Flask webapp for prediction data poision attack

Figure 2. Model Flow Diagram

## V. EVALUATION METRICS

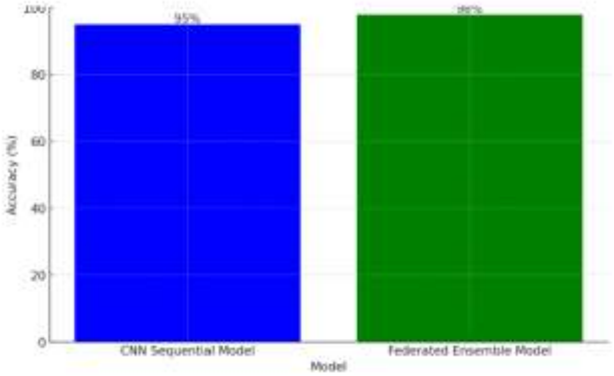
Comparative analysis on different algorithms

- **CNN Algorithm** : Accuracy 95 percent
- **Federated Ensemble Model** : Accuracy 98 percent

| Algorithm | Accuracy |
|---|---|
| **CNN Algorithm** | Accuracy 95 percent |
| **Federated Ensemble Model** | Accuracy 98 percent |

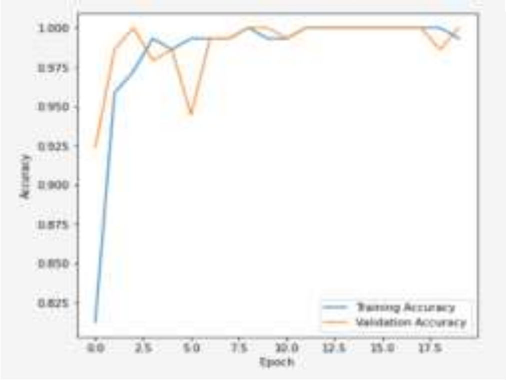
**COMPARISION GRAPH:**

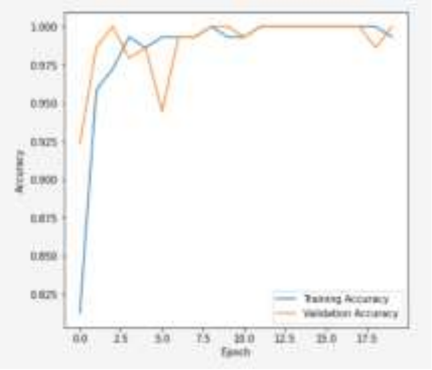
- Comparison of algorithms (CNN sequential model, Federated Ensemble model)



Above bar graph shows accuracy comparison of various algorithms in which CNN sequential model shows 95 percent accuracy and Federated Ensemble model shows 98 percent accuracy.

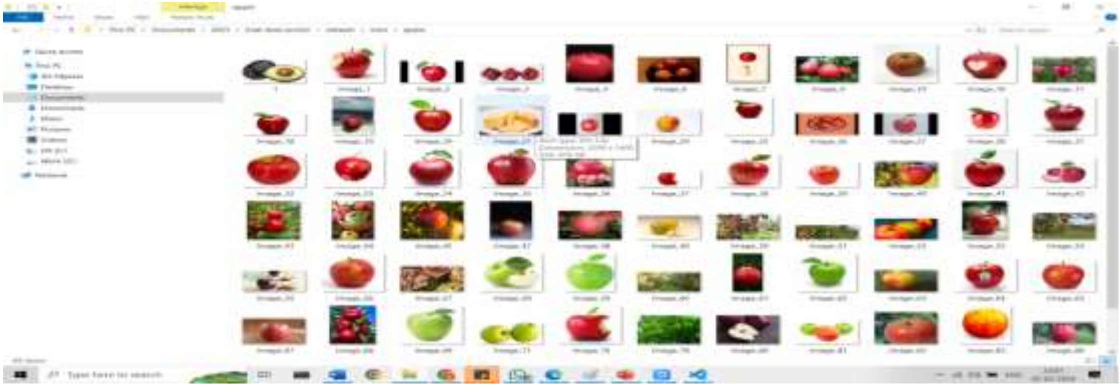
**Accuracy Graphs:**



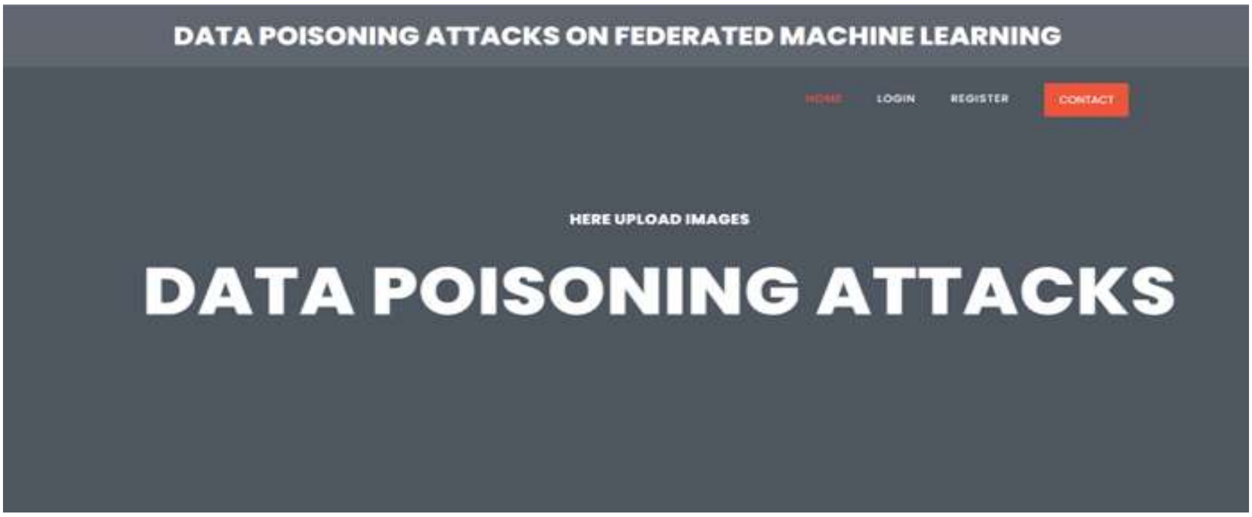


**RESULTS:**

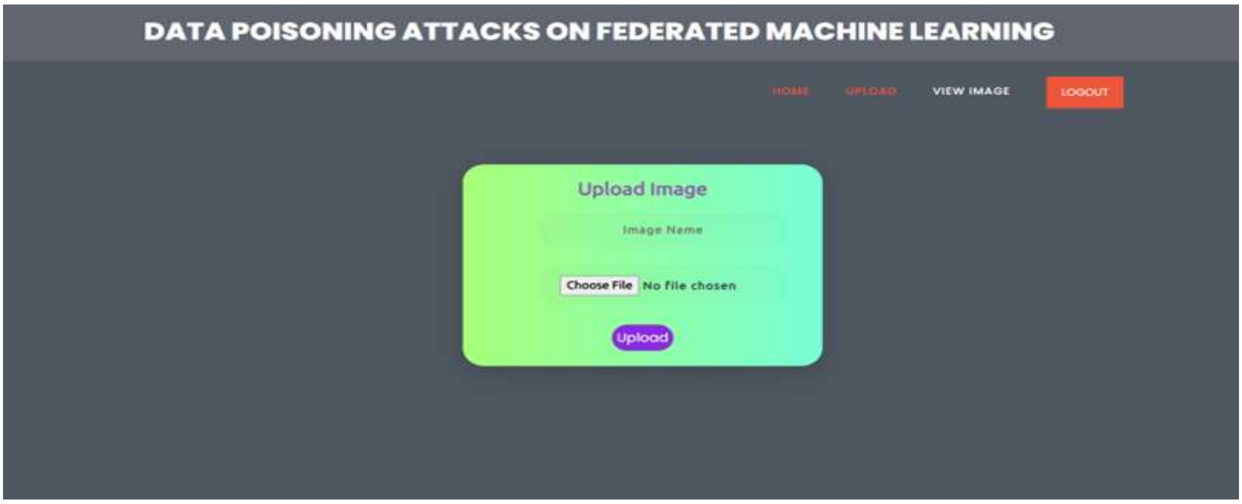
**DATASET:**



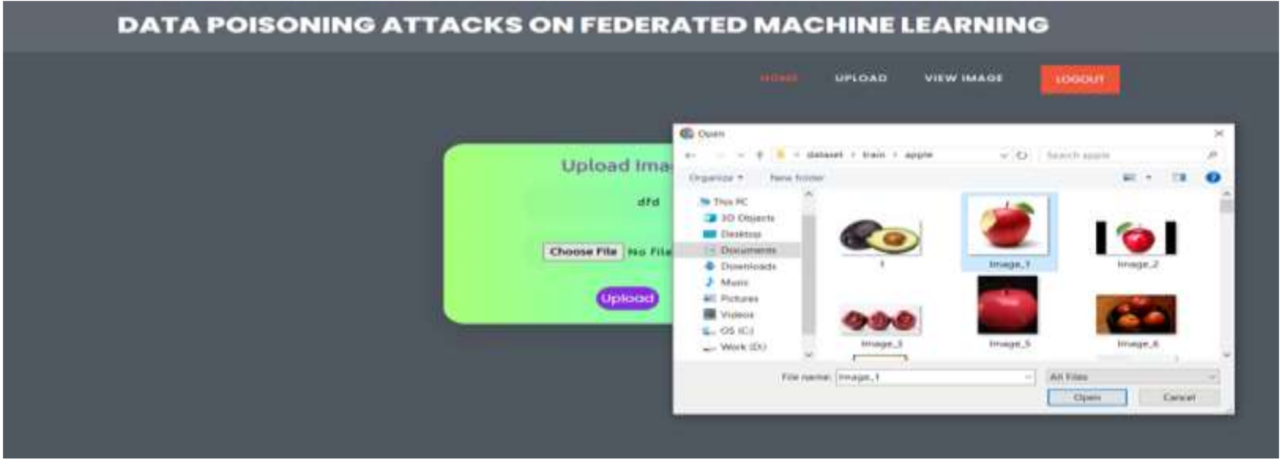
**Home Page:**



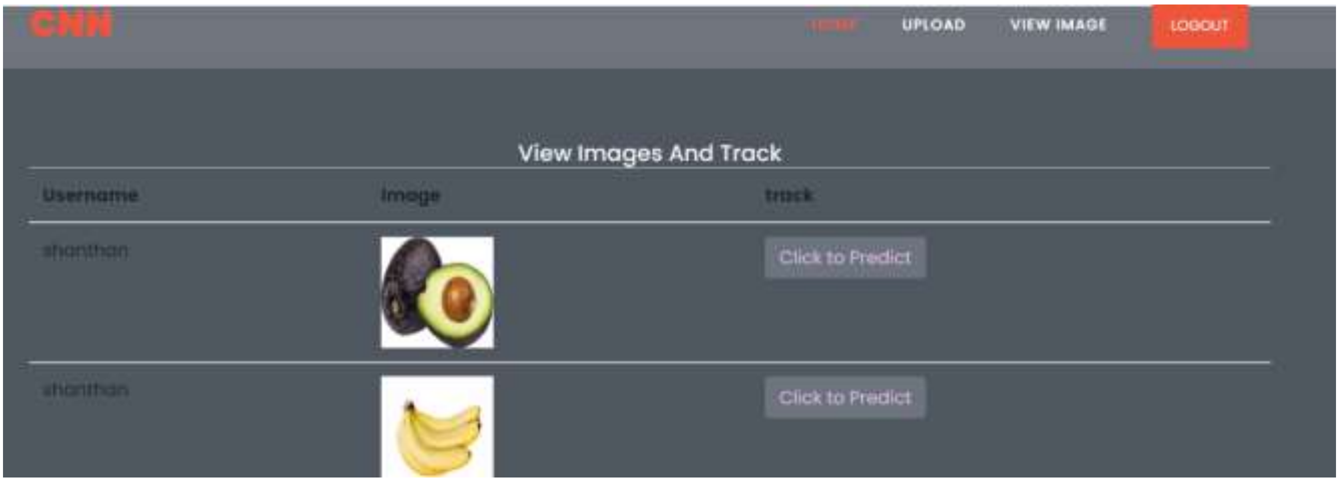
**Input Form:**



**Upload Data:**



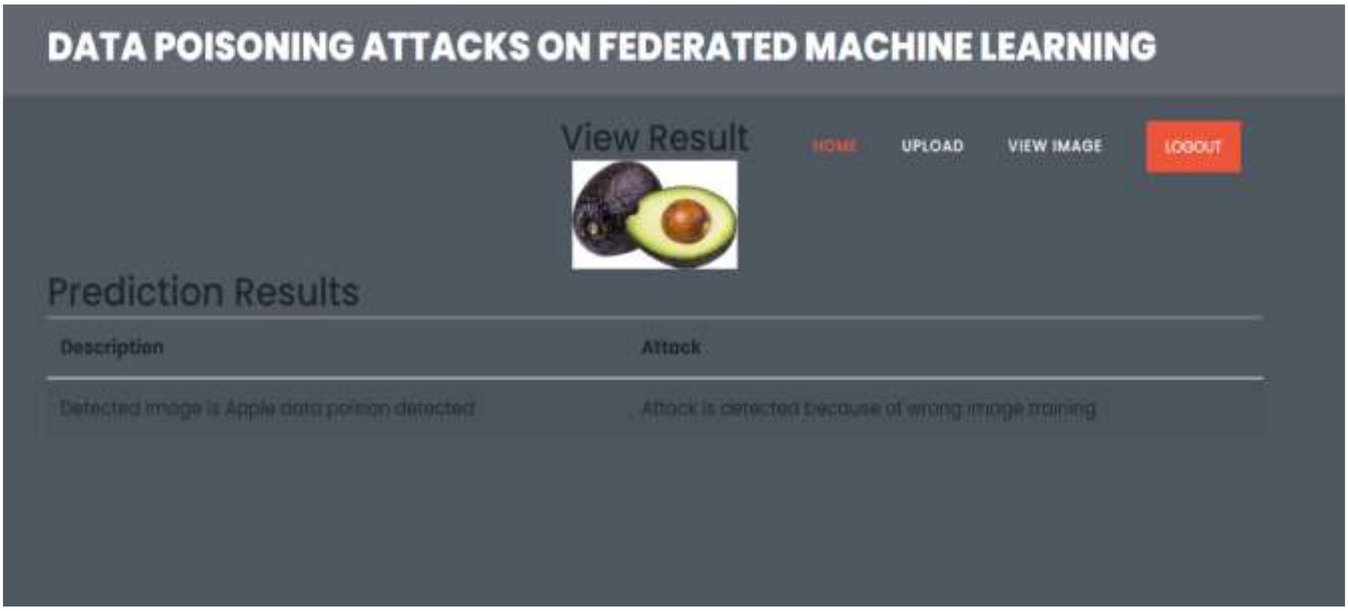
**Result:**



View Images And Track

| Username | Image | track |
|---|---|---|
| shanthan | | Click to Predict |
| shanthan | | Click to Predict |

Result:



Prediction Results

| Description | Attack |
|---|---|
| Detected image is Apple data poision detected | Attack is detected because of wrong image training |

## VI. CONCLUSION

In this study, we proposed a lightweight CNN-based federated ensemble model by combining resent, inceptionv3 and efficient net to classify the two images dataset with a poison data to train images dataset of fruit images. According to the results of the experiments, the proposed CNN federated ensemble model achieves remarkable results in attacks detected classification and can also be used as a feature extraction tool for the traditional Deep learning classifiers. Thus, the proposed CNN ensemble model can be used as an assistance tool for data poisoning attacks in the field to detect wrong training images and bypass the manual process that leads to inaccurate and time-consuming results.

## VII. FUTURE SCOPE

In future work, optimization techniques can be used to obtain optimized values for the hyperparameters of the proposed CNN ensemble model. The proposed model can also be used for predicting other types of problems. Since, the proposed model belongs to the family of low-scale deep learning methods in terms of the number of layers, parameters, and depth. Therefore, a study on using the proposed model in the Industrial Internet of Things (IIoT) domain for classification purposes can be explored.

## VIII. REFERENCES

[[1] Scott Alfeld, Xiaojin Zhu, and Paul Barford. Data poisoning attacks against autoregressive models. In Thirtieth AAAI Conference on Artificial Intelligence, 2016.

[2] Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How to backdoor federated learning. arXiv preprint arXiv:1807.00459, 2018.

[3] Marco Barreno, Blaine Nelson, Anthony D Joseph, and J Doug Tygar. The security of machine learning. Machine Learning, 81(2):121–148, 2010.

[4] Marco Barreno, Blaine Nelson, Russell Sears, Anthony D Joseph, and J Doug Tygar. Can machine learning be secure? In Proceedings of the 2006 ACM Symposium on Information, computer and communications security, pages 16–25. ACM, 2006.

[5] Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. arXiv preprint arXiv:1206.6389, 2012.

[6] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Return of the devil in the details: Delving deep into convolutional nets. arXiv preprint arXiv:1405.3531, 2014.

[7] Ingrid Daubechies, Massimo Fornasier, and Ignace Loris. Accelerated projected gradient method for linear inverse problems with sparsity constraints. journal of fourier analysis and applications, 14(5-6):764 792, 2008

. [8] Adrià Gascón, Phillipp Schoppmann, and Mariana Raykova. Secure linear regression on vertically partitioned datasets.

[9] Shuguo Han, Wee Keong Ng, Li Wan, and Vincent CS Lee. Privacy preserving gradient-descent methods. IEEE Transactions on Knowledge and Data Engineering, 22(6):884–899, 2009.

[10] Ling Huang, Anthony D Joseph, Blaine Nelson, Benjamin IP Rubinstein, and JD Tygar. Adversarial machine learning. In Proceedings of the 4th ACM workshop on Security and artificial intelligence, pages 43–58. ACM, 2011.

[11] Jakub Koneˇcn`y, Brendan McMahan, and Daniel Ramage. Federated optimization: Distributed optimization beyond the datacenter. arXiv preprint arXiv:1511.03575, 2015.

[12] Jakub Koneˇcn` y, H Brendan McMahan, Daniel Ramage, and Peter Richtárik. Federated optimization: Distributed machine learning for on-device intelligence. arXiv preprint arXiv:1610.02527, 2016.

[13] Bo Li, Yining Wang, Aarti Singh, and Yevgeniy Vorobeychik. Data poisoning attacks on factorization-based collaborative filtering. In Advances in neural information processing systems, pages 1885–1893, 2016