CLASSIFICATION OF PHISHING WEBSITES BASED ON MACHINE LEARNING TECHNIQUES

1600003764 – Safa ORHAN
1600003762 – Eyüp USTA

Istanbul Kültür University
Advisor: Assis. Prof. Dr. Öznur ŞENGEL

INTRODUCTION

Phishing, a continuously growing cyber threat, aims to obtain innocent users' credentials by deceiving them via presenting fake web pages which mimic their legitimate targets.

PROBLEM STATEMENT

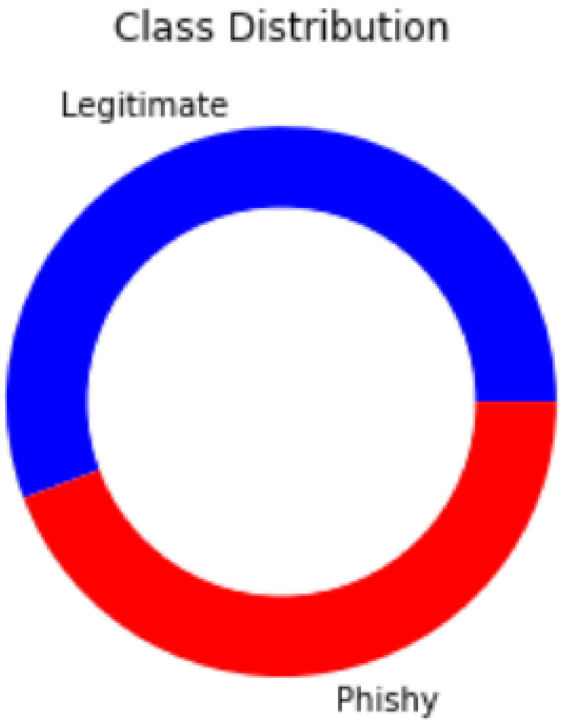
- Total of 5.16 billion people use internet worldwide which is more than 65% of total world population recorded in 2021 by Internet world stats.
- The September 2017 Webroot Data estimates that approximately 1.3 million phishing websites are created on a monthly basis.

WHO	WHAT
WHERE	WHY

OVERVIEW

- Dataset
- Comparison Methods
- Machine Learning Algorithms
- Deep Learning Algorithms
- Comparison of Methods
- Experimental Results
- Conclusions

1. DATASET [1]



- 30 Features
- 11055 Different Samples
- 4 Categories

1 = Legitimate

0 = Suspicious

-1 = Phishing

Category	Feature	Values
Address Bar	Using the IP Address	-1, 1
	Long URL to Hide the Suspicious Part	-1, 0, 1
	Links in <Meta>, <Script> and <Link> tags	-1, 1
	URL's having "@" Symbol	-1, 1
	Redirecting using "//"	-1, 1
	Adding Prefix or Suffix Separated by (-) to the Domain	-1, 1
	Sub Domain and Multi Sub Domains	-1, 0, 1
	HTTPS	-1, 0, 1
	Domain Registration Length	-1, 1
	Favicon	-1, 1
	Using Non-Standard Port	-1, 1
	The Existence of "HTTPS" Token in the Domain Part of the URL	-1, 1
Abnormal	Request URL	-1, 0, 1
	URL of Anchor	-1, 0, 1
	Links in <Meta>, <Script> and <Link> tags	-1, 0, 1
	Server Form Handler (SFH)	-1, 0, 1
	Submitting Information to Email	-1, 1
	Abnormal URL	-1, 1
HTML and JavaScript	Website Forwarding	-1, 0, 1
	Status Bar Customization	-1, 1
	Disabling Right Click	-1, 1
	Using Pop-up Window	-1, 1
	IFrame Redirection	-1, 1
Domain	Age of Domain	-1, 1
	DNS Record	-1, 1
	Website Traffic	-1, 0, 1
	PageRank	-1, 1
	Google Index	-1, 1
	Number of Links Pointing to Page	-1, 1
	Statistical-Reports Based Feature	-1, 1

2. COMPARISON METHODS

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

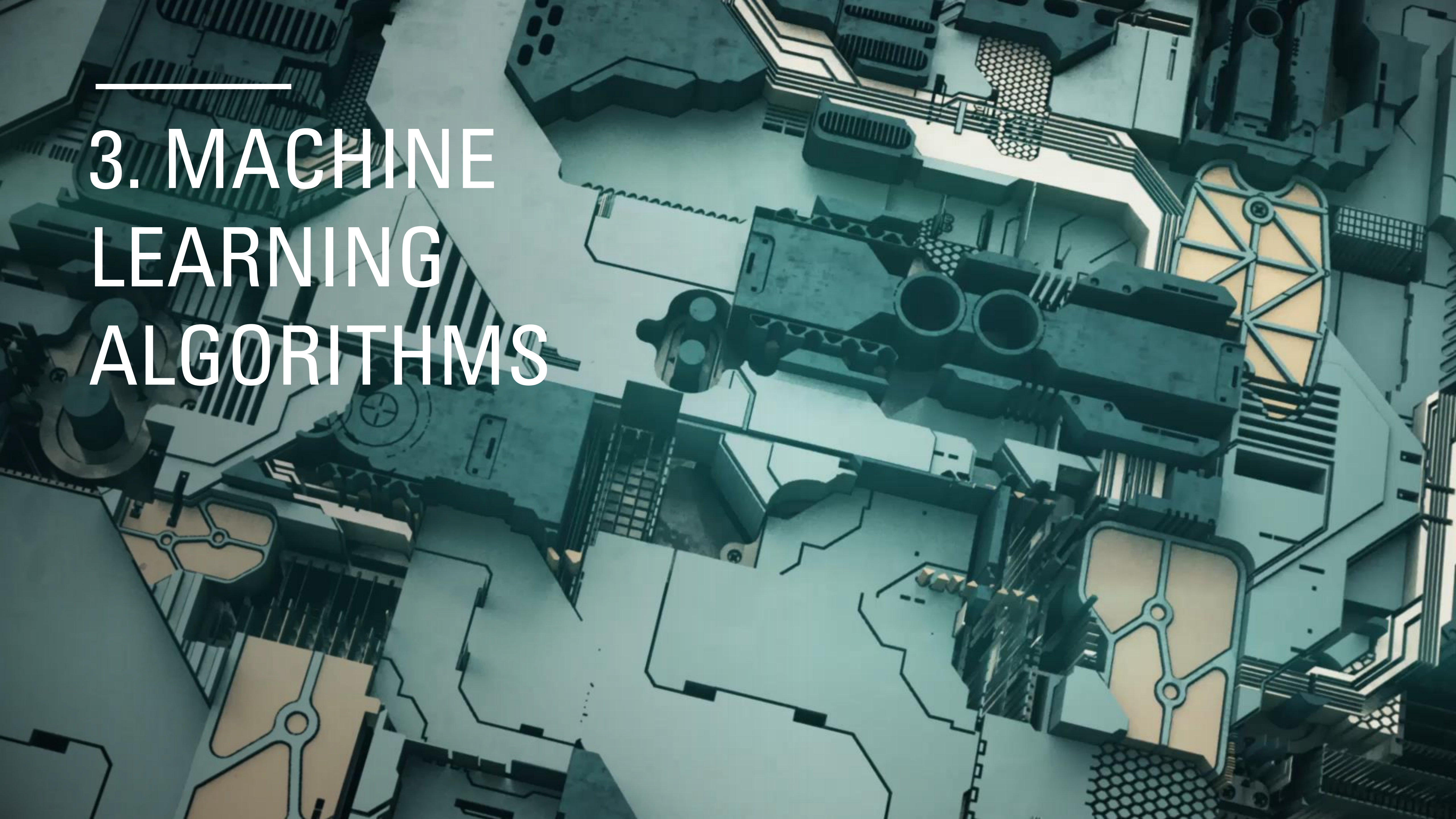
$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

1. Accuracy
2. Precision
3. Recall
4. F-measure (F1-score)

3. MACHINE LEARNING ALGORITHMS



3.1. SUPPORT VECTOR MACHINE (SVM)

Kernels: linear, sigmoid, polynomial, RBF

C: 1, 2, 3, 4, 5

Degree: 1, 2, 3, 4, 5, 6

Gamma: scale, auto

Decision function shape (DFS): ovo, ovr

Test Size: 25%

Model: 2400 Different Combinations

Average Accuracy: 92.42%

3.1. SUPPORT VECTOR MACHINE (SVM)

Best Model:

Kernel: RBF

C: 1

Degree: 2

Gamma: scale

DFS: ovo

Accuracy: 98.37%

Worst Model:

Kernel: sigmoid

C: 3

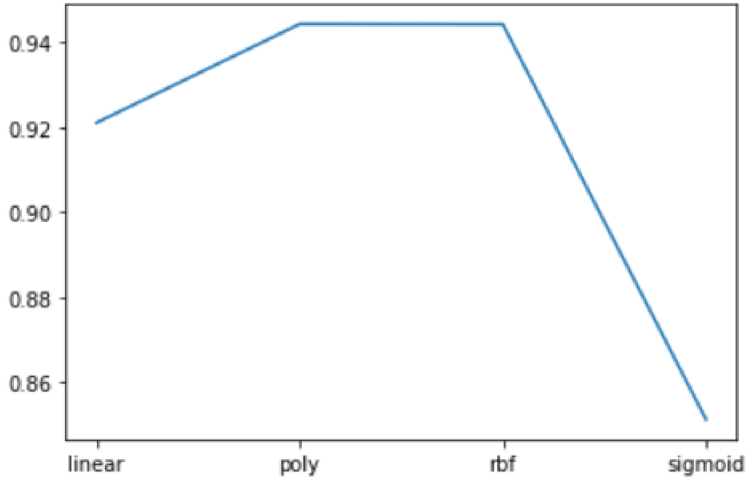
Degree: 1

Gamma: scale

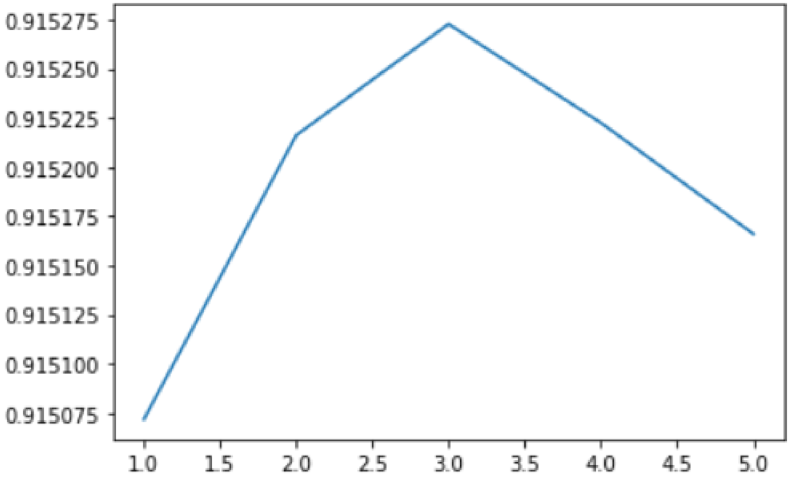
DFS: ovo

Accuracy: 83.78%

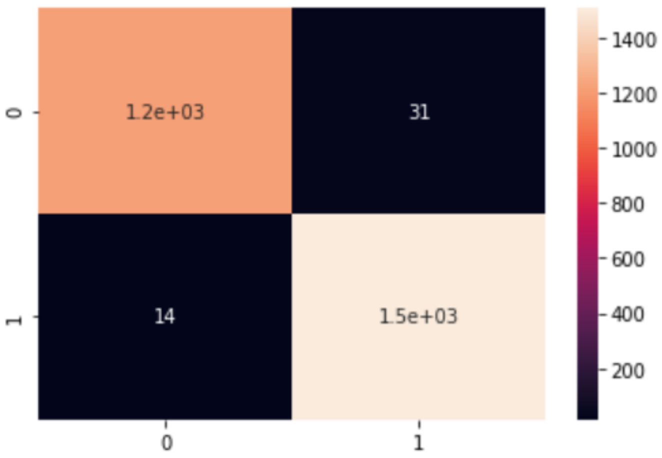
3.1. SUPPORT VECTOR MACHINE (SVM)



Average accuracy values according to kernels



Average accuracy values according to C values



SVM’s Confusion Matrix

3.2. LINEAR SUPPORT VECTOR MACHINE

Losses: hinge, squared hinge

Penalty: L2

C: 1, 2, 3, 4, 5

Multi class: ovr, crammer singer

Test Size: 15%

Model: 100 Different Combinations

Average Accuracy: 93.11%

3.2. LINEAR SUPPORT VECTOR MACHINE

Best Model:

Losses: squared hinge

Penalty: L2

C: 1

Multi class: ovr

Accuracy: 93.97%

Worst Model:

Losses: hinge

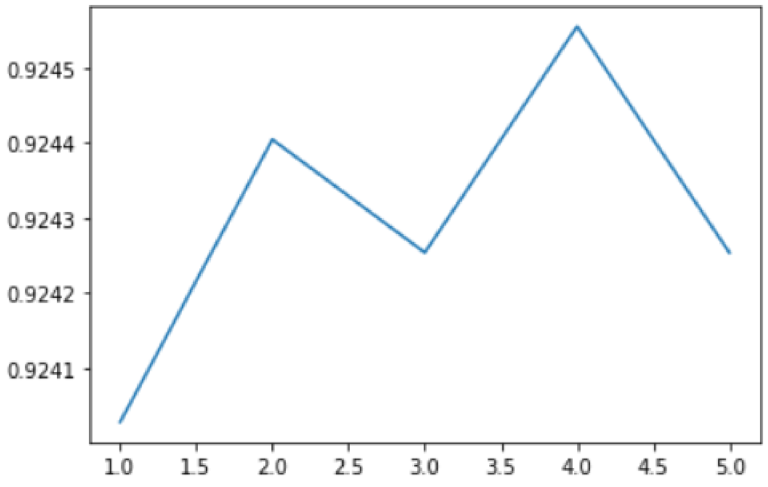
Penalty: L2

C: 4

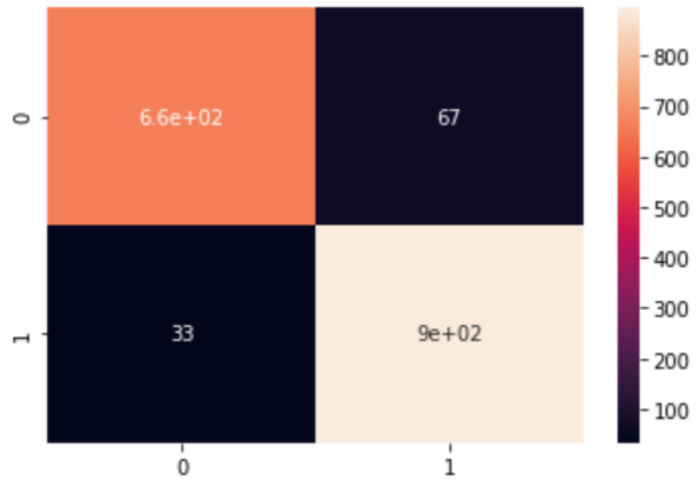
Multi class: ovr

Accuracy: 93.03%

3.2. LINEAR SUPPORT VECTOR MACHINE



Average accuracy values according to C values



Linear SVM's Confusion Matrix

3.3. K-NEAREST NEIGHBORS (KNN)

K: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10

Algorithm: kd tree, ball tree, brute, auto

Weights: uniform, distance

Metric: chebyshev, manhattan, minkowski, euclidean

Test Size: 25%

Model: 1440 Different Combinations

Average Accuracy: 95.99%

3.3. K-NEAREST NEIGHBORS (KNN)

Best Model:

K: 5

Algorithm: auto

Weights: distance

Metric: manhattan

Accuracy: 98.91%

Worst Model:

K: 8

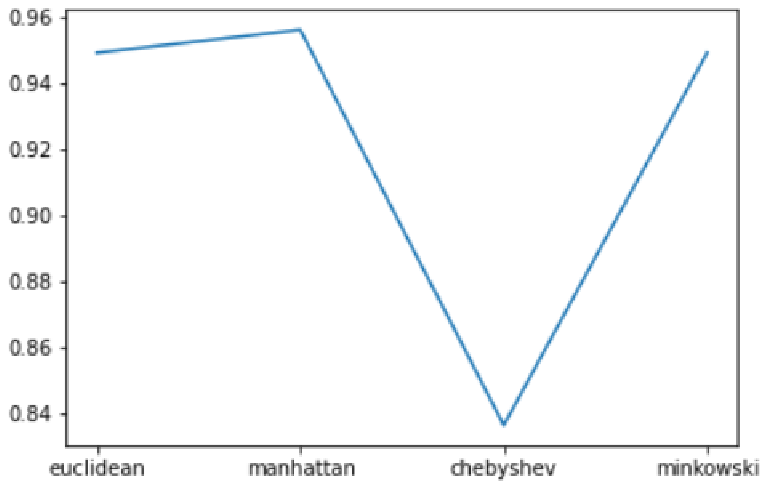
Algorithm: auto

Weights: uniform

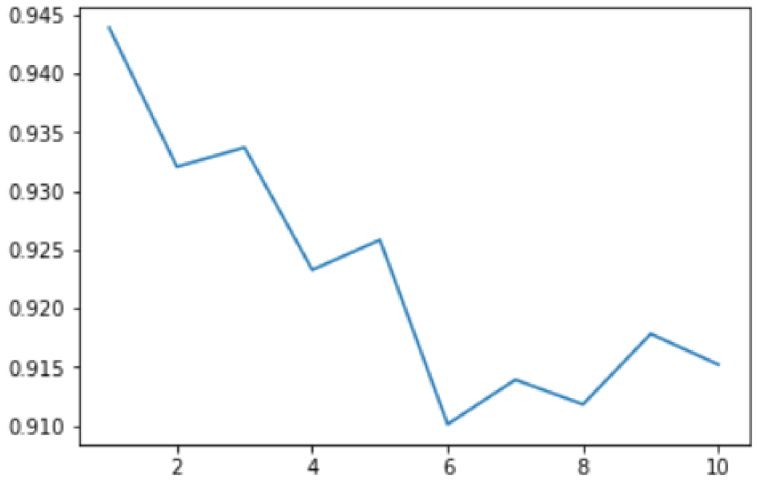
Metric: chebyshev

Accuracy: 78.17%

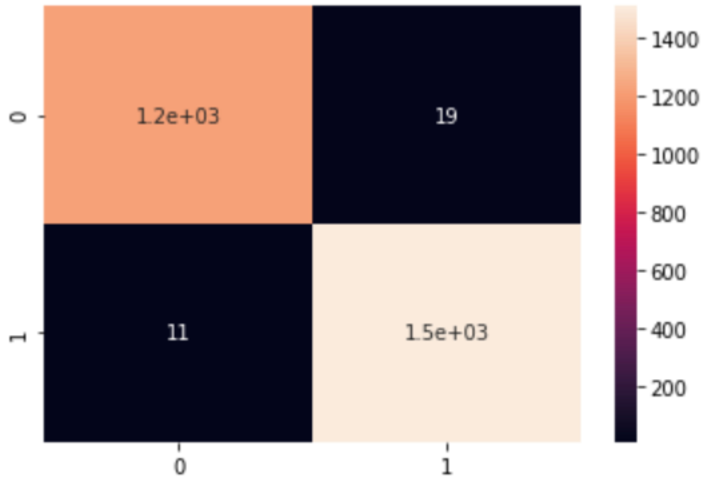
3.3. K-NEAREST NEIGHBORS (KNN)



Average accuracy values according to metrics



Average accuracy values according to K values



KNN’s Confusion Matrix

3.4. DECISION TREE

Criterion: gini, entropy

Max features: auto, none, log2, sqrt

Splitter: random, best

Class weight: none, balanced subsample, balanced

Test Size: 25%

Model: 160 Different Combinations

Average Accuracy: 95.64%

3.4. DECISION TREE

Best Model:

Criterion: entropy

Max features: none

Splitter: best

Class weight: none

Accuracy: 98.76%

Worst Model:

Criterion: gini

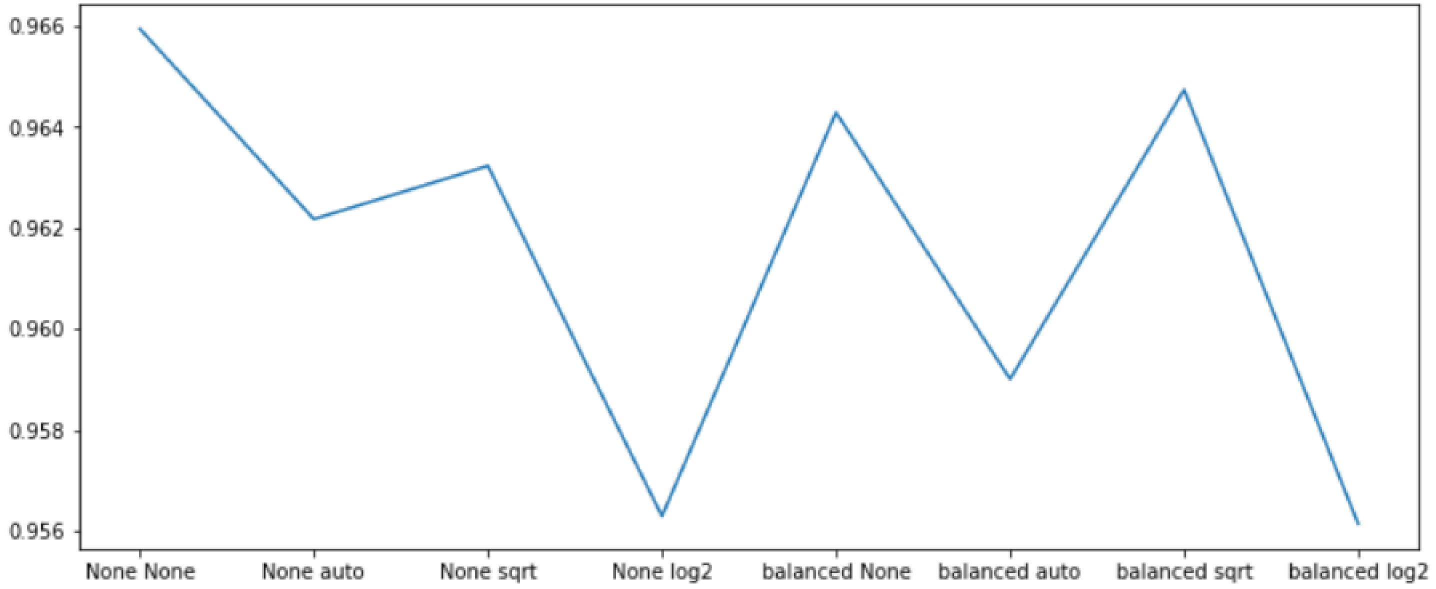
Max features: log2

Splitter: random

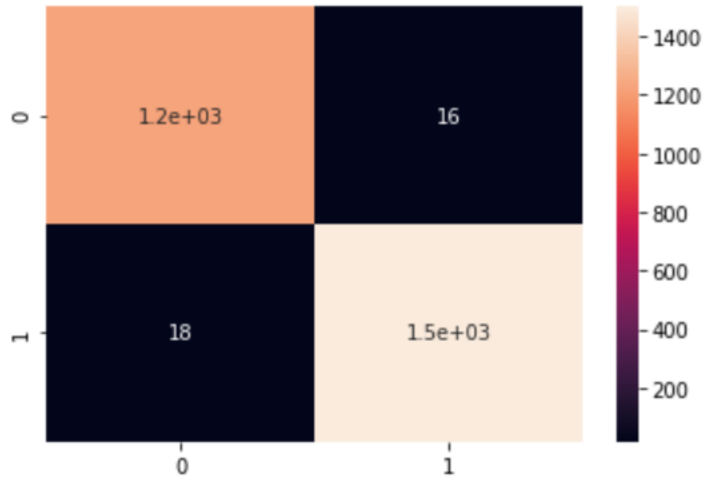
Class weight: none

Accuracy: 95.64%

3.4. DECISION TREE



Average accuracy values according to max features and class weights

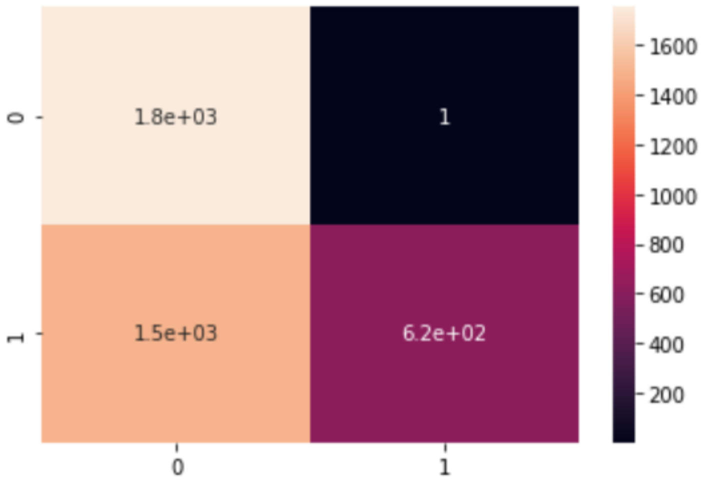


DT's Confusion Matrix

3.5. GAUSSIAN NAIVE BAYES

Test Size: 30%

Accuracy: 61.34%



GNB's Confusion Matrix

3.6. BERNOULLI NAIVE BAYES

Alpha: 0, 1, 2, 3, 4, 5, 7, 9, 11

Binarize: 0, 1, 2, 3, 4, 5, 7, 9, 11

Fit prior: true, false

Test Size: 15%

Model: 810 Different Combinations

Average Accuracy: 60.36%

3.6. BERNOULLI NAIVE BAYES

Best Model:

Alpha: 0

Binarize: 0

Fit prior: false

Accuracy: 91.86%

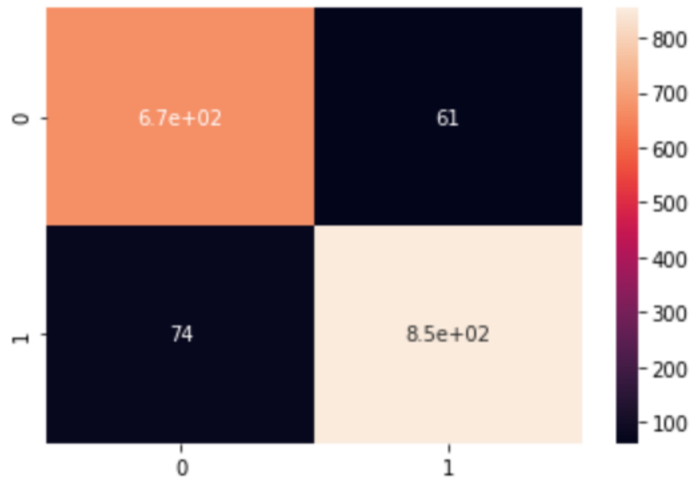
Worst Model:

Alpha: 0

Binarize: 1

Fit prior: true

Accuracy: 56.01%



BNB's Confusion Matrix

3.7. RANDOM FOREST

Max features: auto, none, log2, sqrt

Class weight: none, balanced subsample, balanced

Warm start: true, false

Criterion: entropy, gini

Test Size: 25%

Model: 240 Different Combinations

Average Accuracy: 97%

3.7. RANDOM FOREST

Best Model:

Max features: auto

Class weight: balanced subsample

Warm start: true

Criterion: gini

Accuracy: 99.09%

Worst Model:

Max features: none

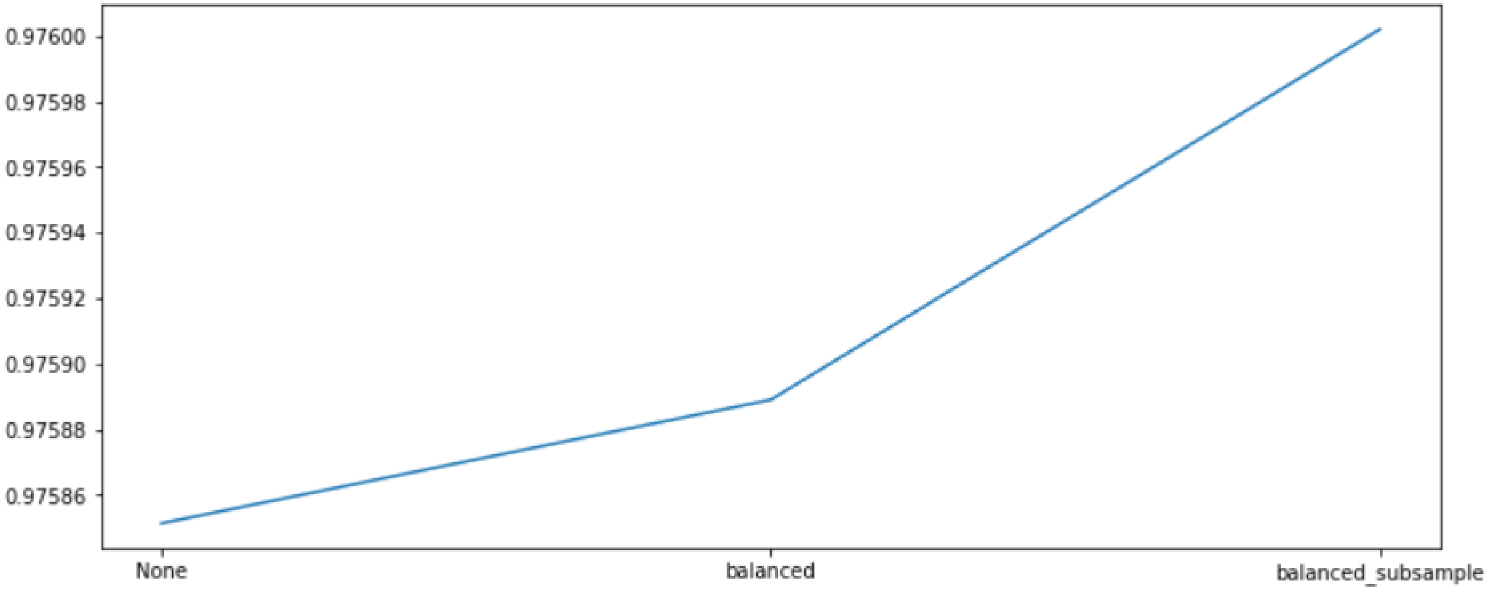
Class weight: balanced

Warm start: true

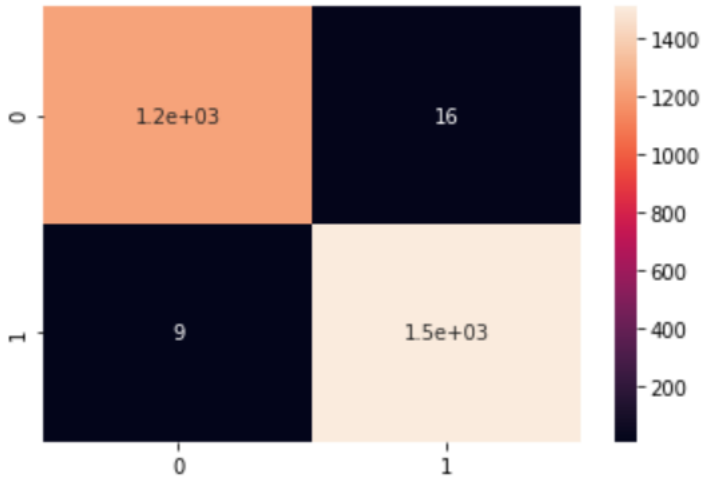
Criterion: entropy

Accuracy: 96.68%

3.7. RANDOM FOREST



Average accuracy values according to class weights



RF's Confusion Matrix

4. DEEP LEARNING ALGORITHMS

4.1. CONVOLUTIONAL NEURAL NETWORKS (CNN)

Epoch: 150

Test Size: 30%

Model: 320 Different Combinations

Activation Functions	Optimizers	Loss Functions
Softplus	SGD	Binary Cross Entropy
Softsign	Rmsprop	Categorical Cross Entropy
Selu	Adam	Hinge
Elu	Adadelta	Squared Hinge
Exponential	Adagrad	Huber
Tanh	Adamax	
Sigmoid	Nadam	
Relu	FTRL	

4.1. CONVOLUTIONAL NEURAL NETWORKS (CNN)

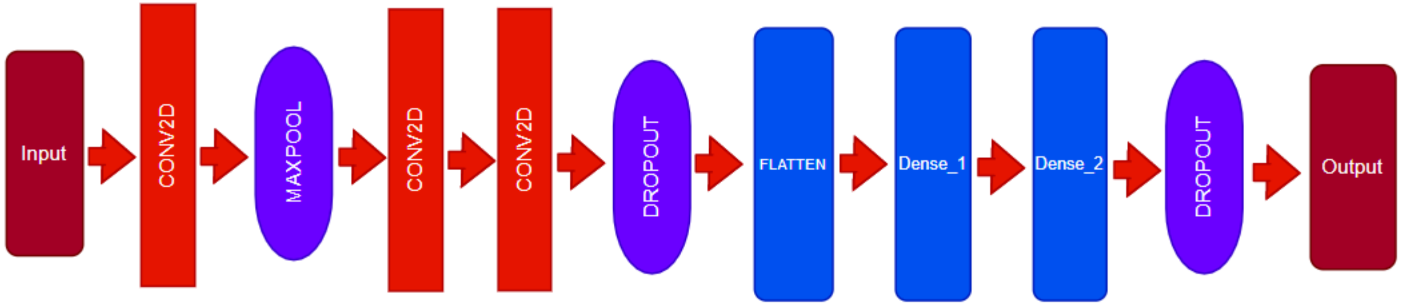
CNN Model 1:

Optimizers: Adamax

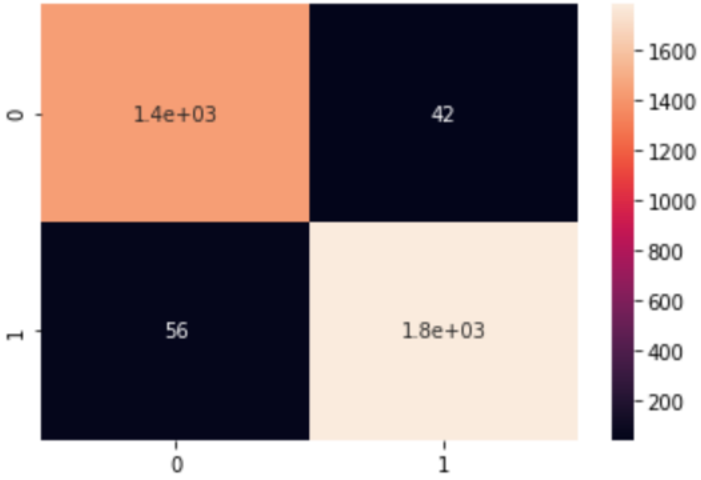
Activation Functions: Tanh

Loss Functions: Huber

Accuracy: 93.19%



CNN Model 1



CNN Model 1’s Confusion Matrix

4.1. CONVOLUTIONAL NEURAL NETWORKS (CNN)

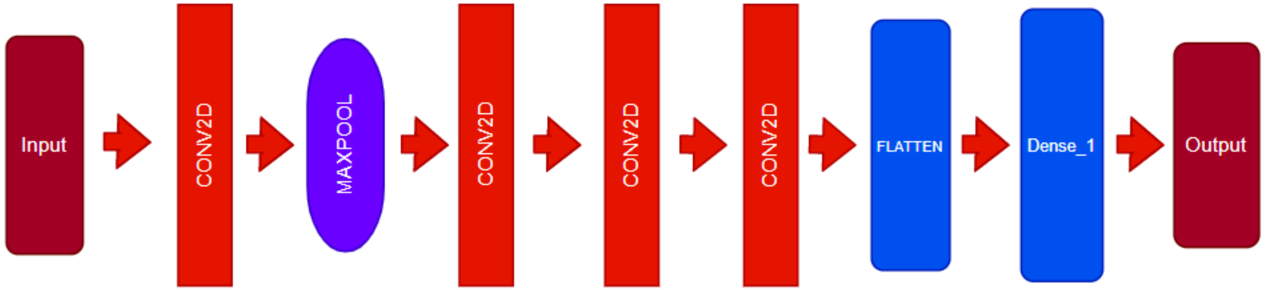
CNN Model 2:

Optimizers: Adam

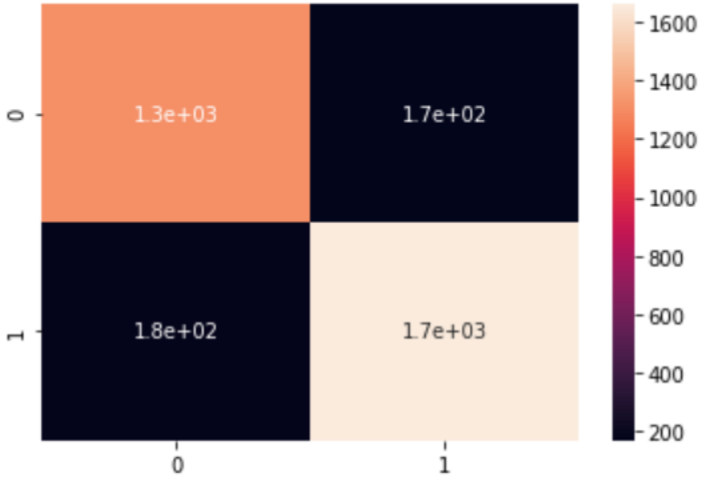
Activation Functions: Tanh

Loss Functions: Binary cross entropy

Accuracy: 92%



CNN Model 2



CNN Model 2’s Confusion Matrix

4.2. RECURRENT NEURAL NETWORKS (RNN)

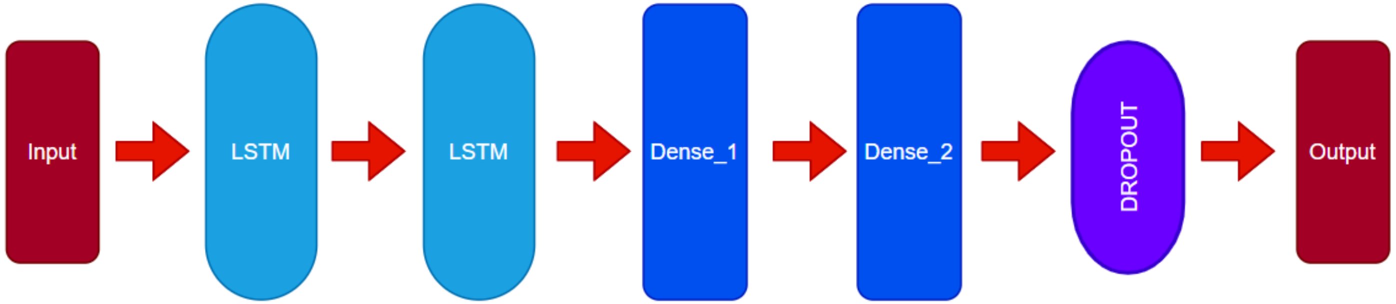
Epoch: 150

Test Size: 30%

Model: 320 Different Combinations

Activation Functions	Optimizers	Loss Functions
Softplus	SGD	Binary Cross Entropy
Softsign	Rmsprop	Categorical Cross Entropy
Selu	Adam	Hinge
Elu	Adadelta	Squared Hinge
Exponential	Adagrad	Huber
Tanh	Adamax	
Sigmoid	Nadam	
Relu	FTRL	

4.2. RECURRENT NEURAL NETWORKS (RNN)



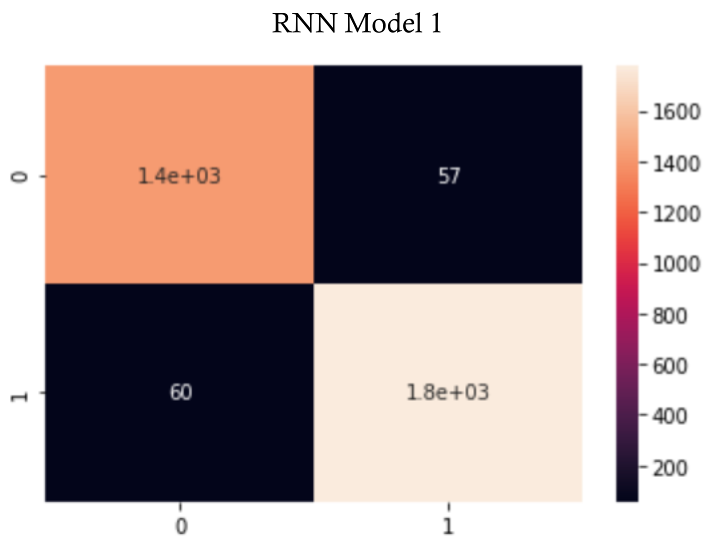
RNN Model 1:

Optimizers: Adamax

Activation Functions: Tanh

Loss Functions: Binary cross entropy

Accuracy: 98.81%



RNN Model 1's Confusion Matrix

4.2. RECURRENT NEURAL NETWORKS (RNN)

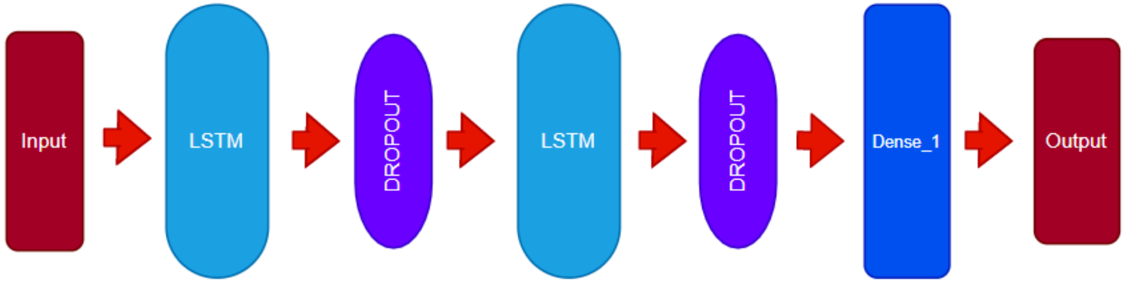
RNN Model 2:

Optimizers: Adamax

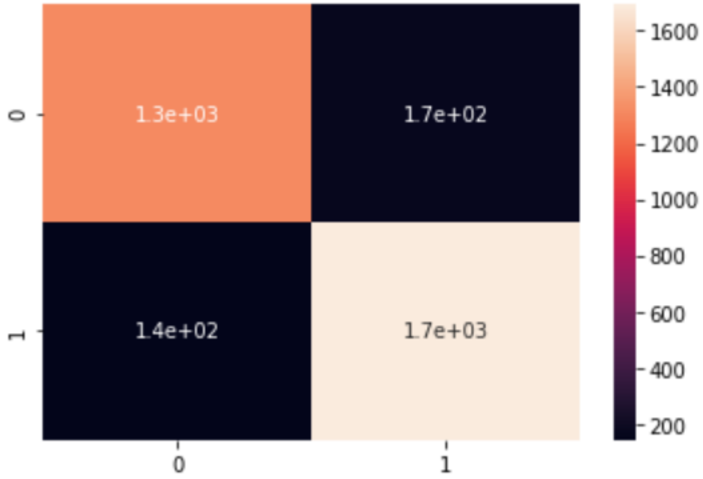
Activation Functions: Sigmoid

Loss Functions: Binary cross entropy

Accuracy: 98.31%



RNN Model 2



RNN Model 2's Confusion Matrix

4.2. RECURRENT NEURAL NETWORKS (RNN)

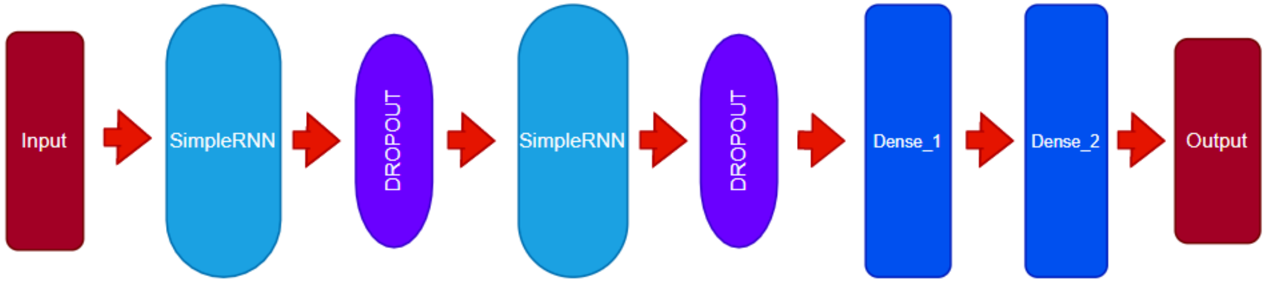
RNN Model 3:

Optimizers: SGD

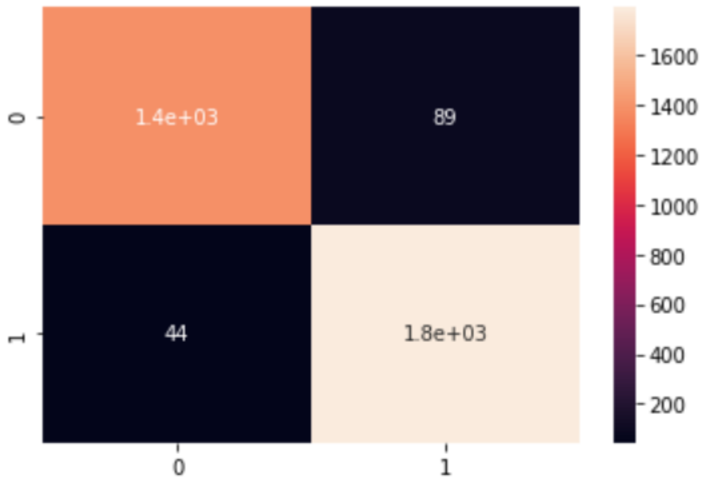
Activation Functions: Sigmoid

Loss Functions: Binary cross entropy

Accuracy: 96.85%

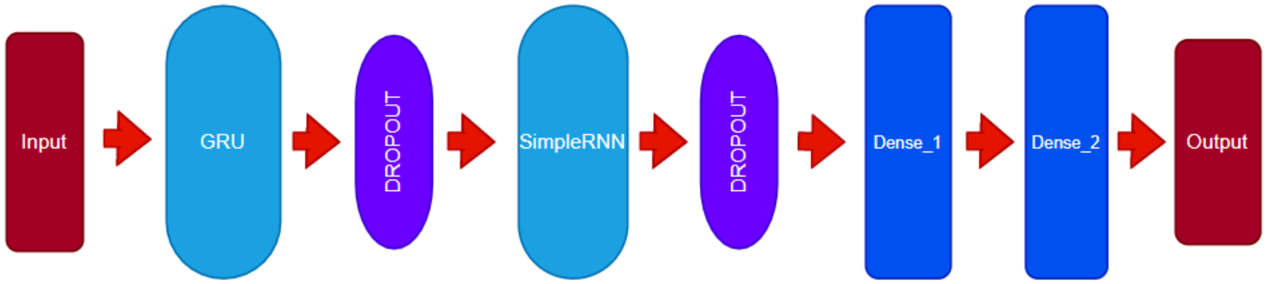


RNN Model 3



RNN Model 3's Confusion Matrix

4.2. RECURRENT NEURAL NETWORKS (RNN)



RNN Model 4

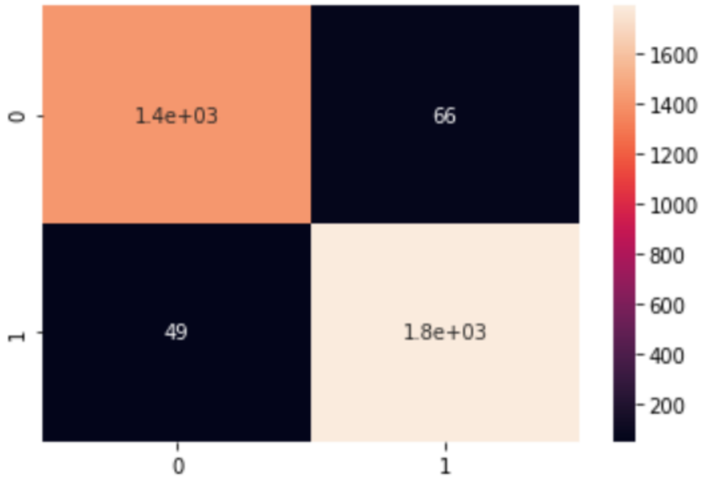
RNN Model 4:

Optimizers: SGD

Activation Functions: Sigmoid

Loss Functions: Binary cross entropy

Accuracy: 96.21%



RNN Model 4's Confusion Matrix

4.3. DEEP NEURAL NETWORKS (DNN)

Epoch: 150

Test Size: 30%

Model: 320 Different Combinations

Activation Functions	Optimizers	Loss Functions
Softplus	SGD	Binary Cross Entropy
Softsign	Rmsprop	Categorical Cross Entropy
Selu	Adam	Hinge
Elu	Adadelta	Squared Hinge
Exponential	Adagrad	Huber
Tanh	Adamax	
Sigmoid	Nadam	
Relu	FTRL	

4.3. DEEP NEURAL NETWORKS (DNN)

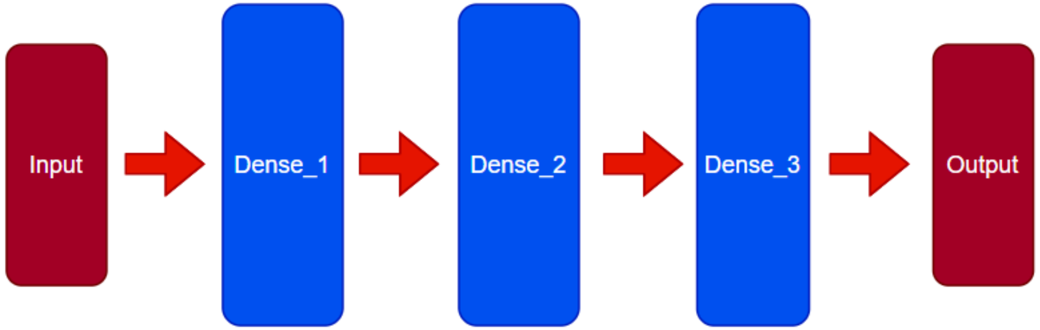
DNN Model 1:

Optimizers: SGD

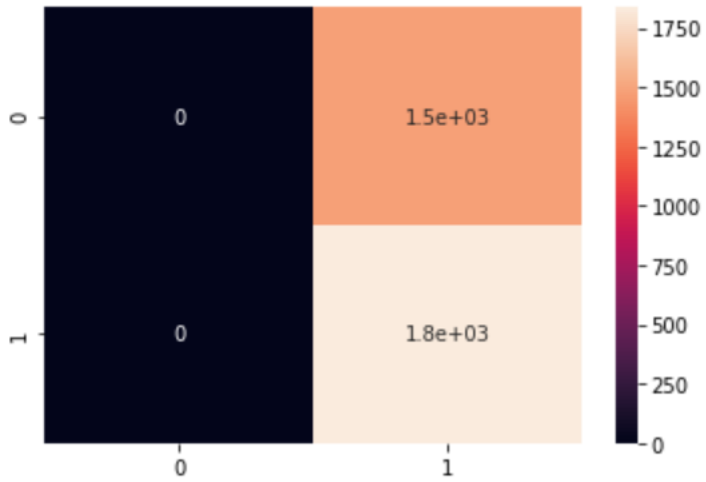
Activation Functions: Sigmoid

Loss Functions: Binary cross entropy

Accuracy: 55.84%



DNN Model 1



DNN Model 1's Confusion Matrix

4.3. DEEP NEURAL NETWORKS (DNN)

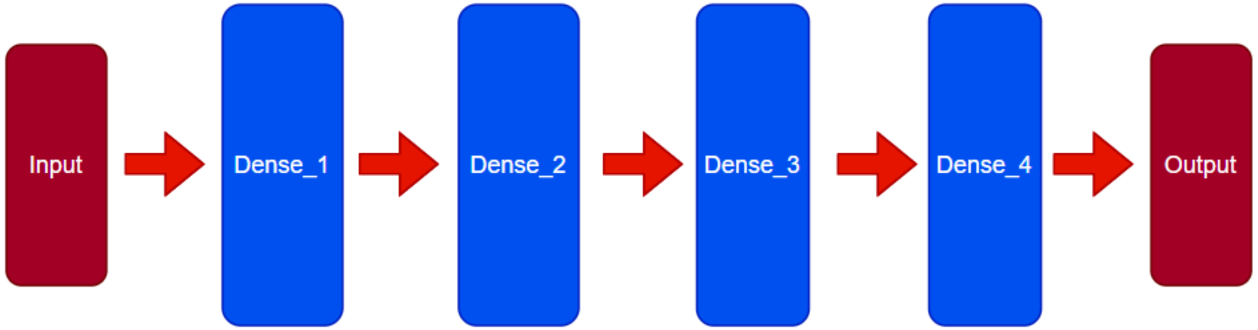
DNN Model 2:

Optimizers: Adamax

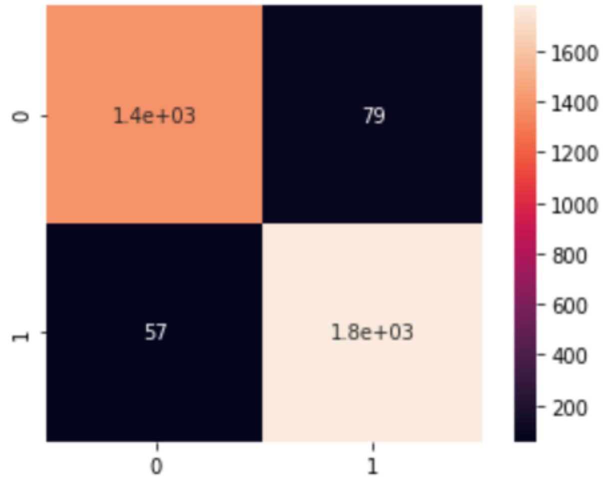
Activation Functions: Sigmoid

Loss Functions: Huber

Accuracy: 98.84%



DNN Model 2



DNN Model 2’s Confusion Matrix

4.4. COMBINED MODELS

Epoch: 150

Test Size: 30%

Activation Functions	Optimizers	Loss Functions
Softplus	SGD	Binary Cross Entropy
Softsign	Rmsprop	Categorical Cross Entropy
Selu	Adam	Hinge
Elu	Adadelta	Squared Hinge
Exponential	Adagrad	Huber
Tanh	Adamax	
Sigmoid	Nadam	
Relu	FTRL	

4.4. COMBINED MODELS

Combined Model 1:

Combination: RNN Model 1 – CNN Model 1

Optimizers: Adamax

Activation Functions: Softplus

Loss Functions: Binary cross entropy

Accuracy: 90.02%

Combined Model 2:

Combination: CNN Model 1 – RNN Model 1

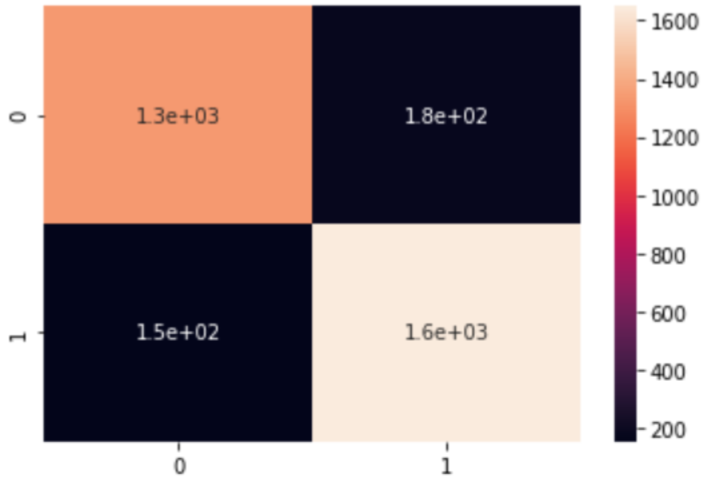
Optimizers: Adamax

Activation Functions: Softplus

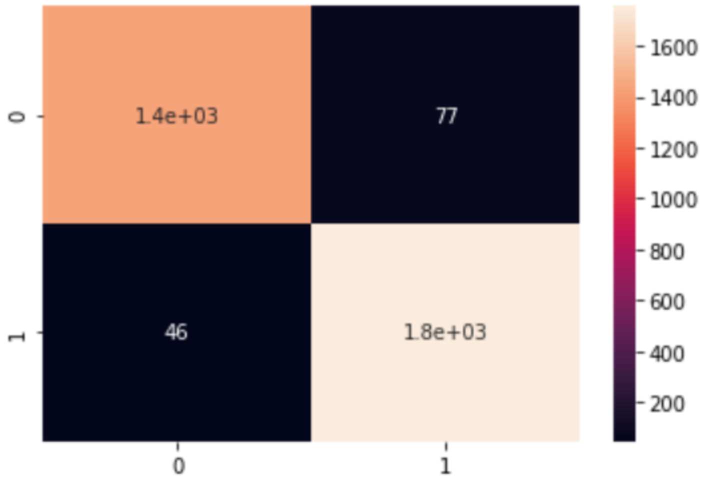
Loss Functions: Binary cross entropy

Accuracy: 96.29%

4.4. COMBINED MODELS



Combined Model 1's Confusion Matrix



Combined Model 2's Confusion Matrix

4.4. COMBINED MODELS

Combined Model 3:

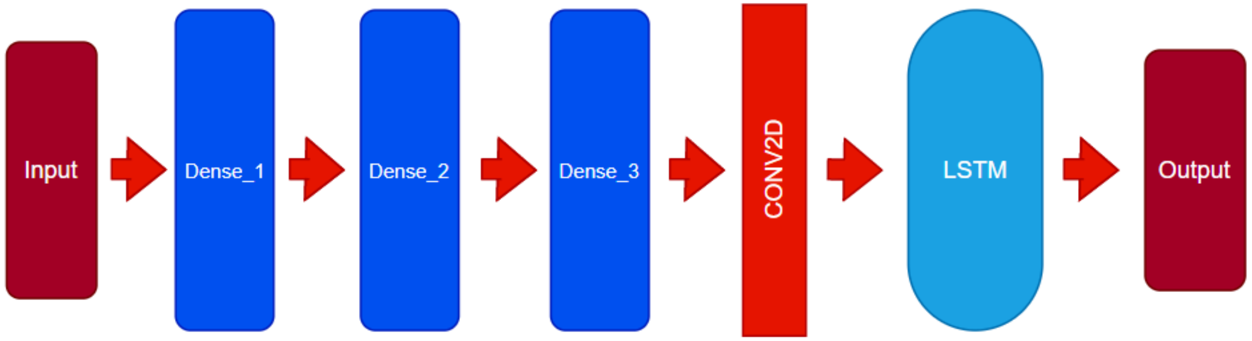
Combination: CNN – LSTM

Optimizers: Adamax

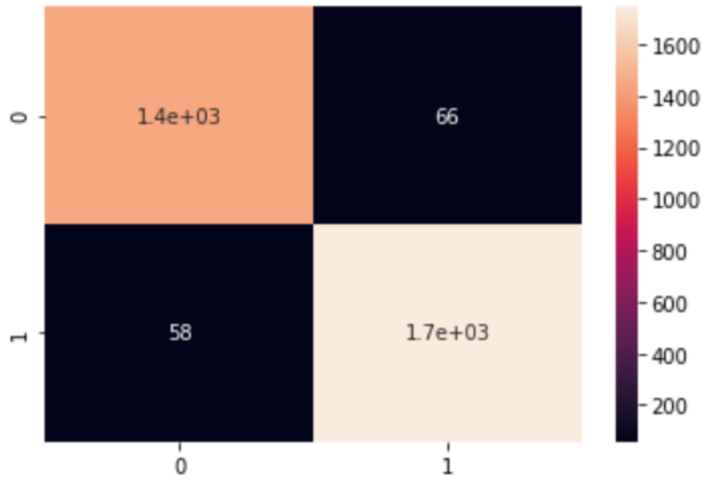
Activation Functions: Softplus

Loss Functions: Binary cross entropy

Accuracy: 96.26%



Combined Model 3



Combined Model 3's Confusion Matrix

4.5. MODEL TESTS WITH DIFFERENT DATASETS

First Dataset [2]:

- 48 Features
- 10000 Different Samples
- From data mendeley

Second Dataset [3]:

- 14 Features
- 11000 Different Samples
- From kaggle

Third Dataset [4]:

- 35 Features
- 13071 Different Samples
- From data mendeley

Fourth Dataset [5]:

- 87 Features
- 11430 Different Samples
- From data mendeley

Fifth Dataset [6]:

- 79 Features
- 15367 Different Samples
- From Canadian Institute for Cybersecurity

4.5. MODEL TESTS WITH DIFFERENT DATASETS

Dataset	DNN Model 2	RNN Model 1	CNN Model 1
First Dataset [2]	95.43%	93.93%	88.87%
Second Dataset [3]	89.45%	99.12%	99.68%
Third Dataset [4]	94.11%	93.19%	92.76%
Fourth Dataset [5]	75.51%	77.34%	77.81%
Fifth Dataset [6]	48.88%	48.88%	48.88%
Our Dataset [1]	98.84%	98.81%	93.19%

5. COMPARISON OF METHODS

Model	Accuracy	Precision	Recall	F1-Score
Our RF Model	99.09%	0.9895	0.9940	0.9917
RNN Model 2	97.27%	0.9769	0.9694	0.9732
Combined Model 3	96.26%	0.9635	0.9678	0.9656
[7]	98.61%	0.9941	0.9857	0.99
[8]	95.60%	0.9733	0.9378	0.9552
[9]	93.28%	0.9327	0.9330	0.9329
Random Forest [10]	98.03%	1.0	0.96	0.98
Logistic Regression [10]	97.7%	1.0	0.96	0.98
Gaussian Naïve Bayes [10]	97.18%	1.0	0.95	0.97
[11]	93%	0.92		0.91
CNN [12]	97.42%	0.9648	0.9723	
LSTM [12]	99.14%	0.9874	0.9891	

6. EXPERIMENTAL RESULTS

Models	Features	Accuracy	Test Size	Epoch
DNN Model 1	Categorical	55.84%	30%	150
DNN Model 1	Categorical	75.76%	15%	10
DNN Model 1	Categorical	75.54%	35%	10
DNN Model 2	Categorical	75.79%	15%	10
DNN Model 2	Categorical	75.40%	35%	10
DNN Model 2	Non-categorical	98.84%	30%	150
RNN Model 1	Non-categorical	98.81%	30%	150
RNN Model 2	Non-categorical	98.31%	30%	150
RNN Model 3	Non-categorical	96.85%	30%	150
RNN Model 4	Non-categorical	96.21%	30%	150
CNN Model 1	Non-categorical	93.19%	30%	150
CNN Model 2	Non-categorical	92%	30%	150
Combined Model 1	Non-categorical	90.02%	30%	150
Combined Model 2	Non-categorical	96.29%	30%	150
Combined Model 3	Non-categorical	96.26%	30%	150

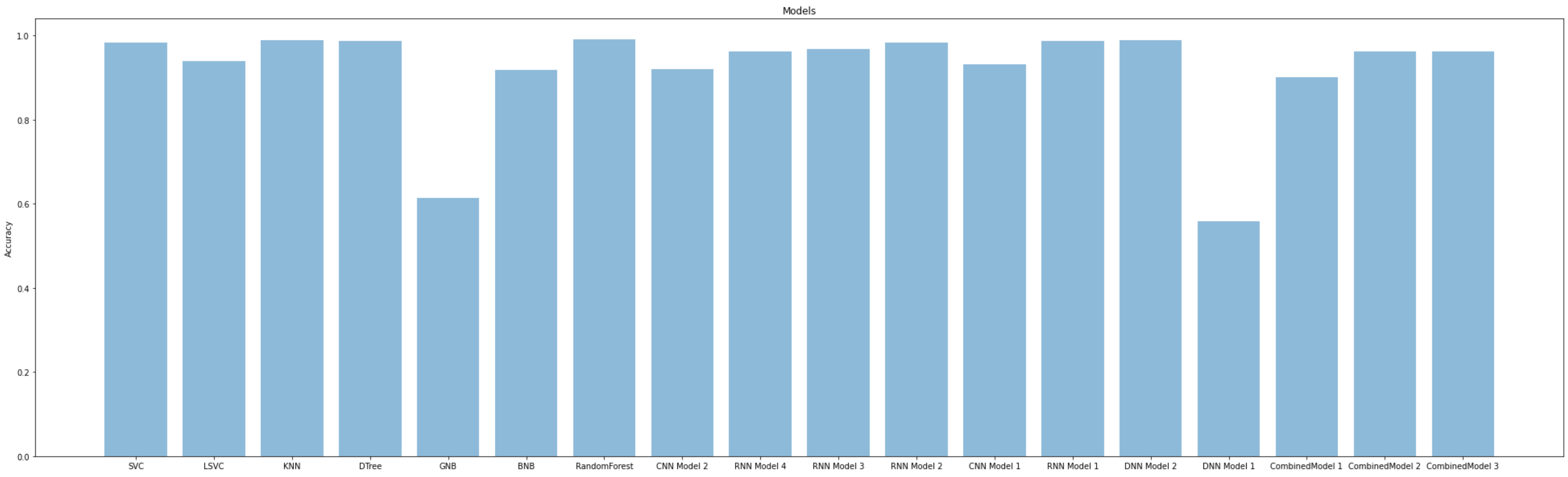
6. EXPERIMENTAL RESULTS

Model	Parameters	Train Accuracy	Test Accuracy	Epoch	Test Size
RNN Model 1	Tanh/Binary Cross Entropy/Adamax	98.81%	95.99%	150	30%
RNN Model 2	Sigmoid/Binary Cross Entropy/Adamax	98.31%	97.20%	150	30%
RNN Model 3	Sigmoid/Binary Cross Entropy/SGD	96.85%	96.47%	150	30%
RNN Model 4	Sigmoid/Binary Cross Entropy/SGD	96.21%	95.90%	150	30%
DNN Model 1	Sigmoid/Binary Cross Entropy/SGD	55.84%	55.35%	150	30%
DNN Model 2	Sigmoid/Adamax/Huber	98.84%	96.53%	150	30%
CNN Model 1	Tanh/Huber/Adamax	93.19%	90.59%	150	30%
CNN Model 2	Tanh/Binary Cross Entropy/Adam	92%	89.63%	150	30%
Combined Model 1	Softplus/Binary Cross Entropy/Adamax	91.23%	90.02%	150	30%
Combined Model 2	Softplus/Binary Cross Entropy/Adamax	98.48%	96.29%	150	30%
Combined Model 3	Softplus/Binary Cross Entropy/Adamax	98.54%	96.26%	150	30%

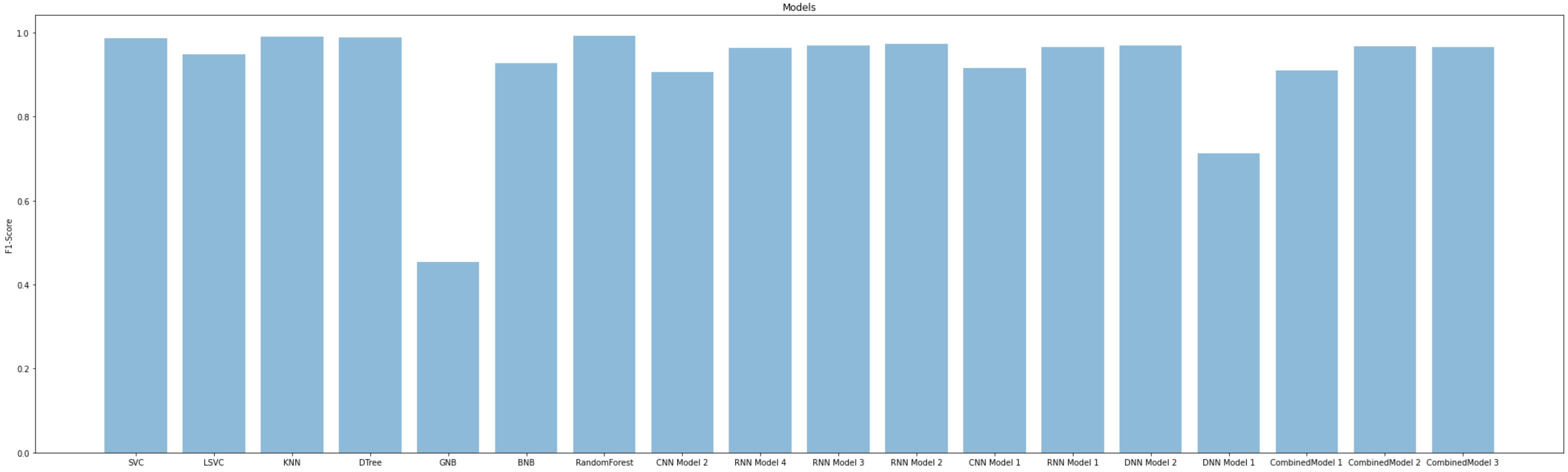
6. EXPERIMENTAL RESULTS

Model	Accuracy	Cross–Validation Accuracy
SVM	98.37%	94.74%
Bernoulli Naïve Bayes	91.86%	90.40%
KNN	98.91%	96.63%
Linear SVM	93.97%	92.65%
Gaussian Naïve Bayes	61.34%	60.38%
Decision Tree	98.76%	96.43%
Random Forest	99.09%	97.24%

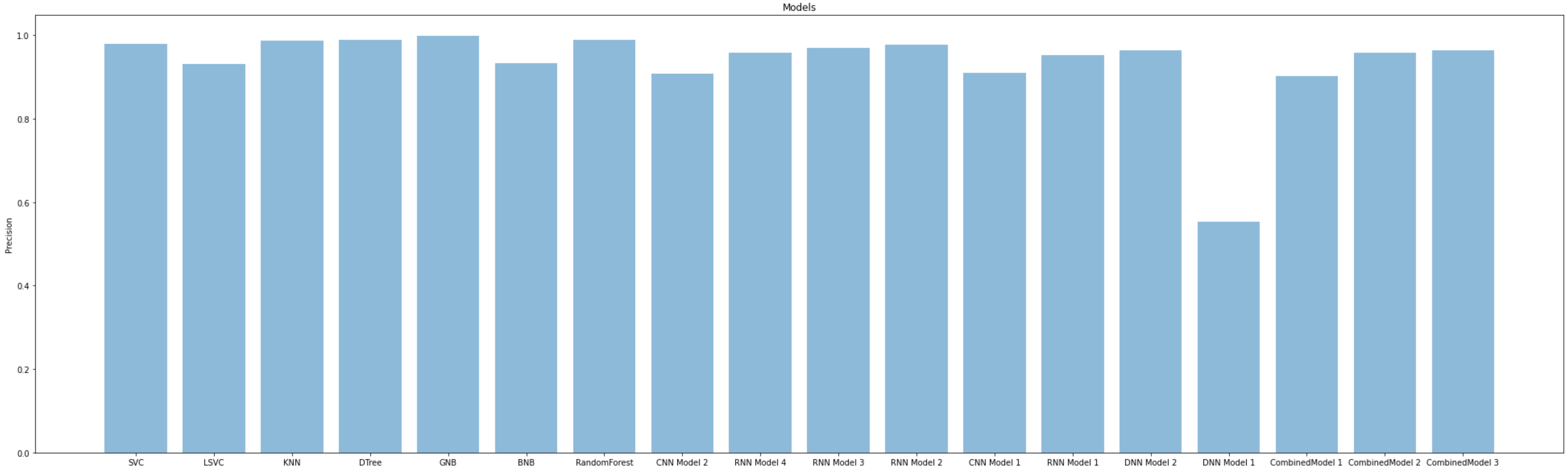
Model	F1 Score	Precision	Recall
SVM	0.9852	0.9798	0.9907
Linear SVM	0.9470	0.9303	0.9644
KNN	0.9901	0.9875	0.9927
Gaussian Naïve Bayes	0.4544	0.9983	0.2941
Decision Tree	0.9888	0.9894	0.9881
Bernoulli Naïve Bayes	0.9267	0.9333	0.9202
Random Forest	0.9917	0.9895	0.9940
CNN Model 2	0.9061	0.9080	0.9041
CNN Model 1	0.9156	0.9092	0.9221
DNN Model 2	0.9688	0.9643	0.9733
DNN Model 1	0.7125	0.5535	1.0
RNN Model 4	0.9631	0.9574	0.9689
RNN Model 3	0.9681	0.9689	0.9673
RNN Model 2	0.9732	0.9769	0.9694
RNN Model 1	0.9642	0.9526	0.9760
Combined Model 1	0.9087	0.9025	0.9150
Combined Model 2	0.9661	0.9579	0.9744
Combined Model 3	0.9656	0.9635	0.9678



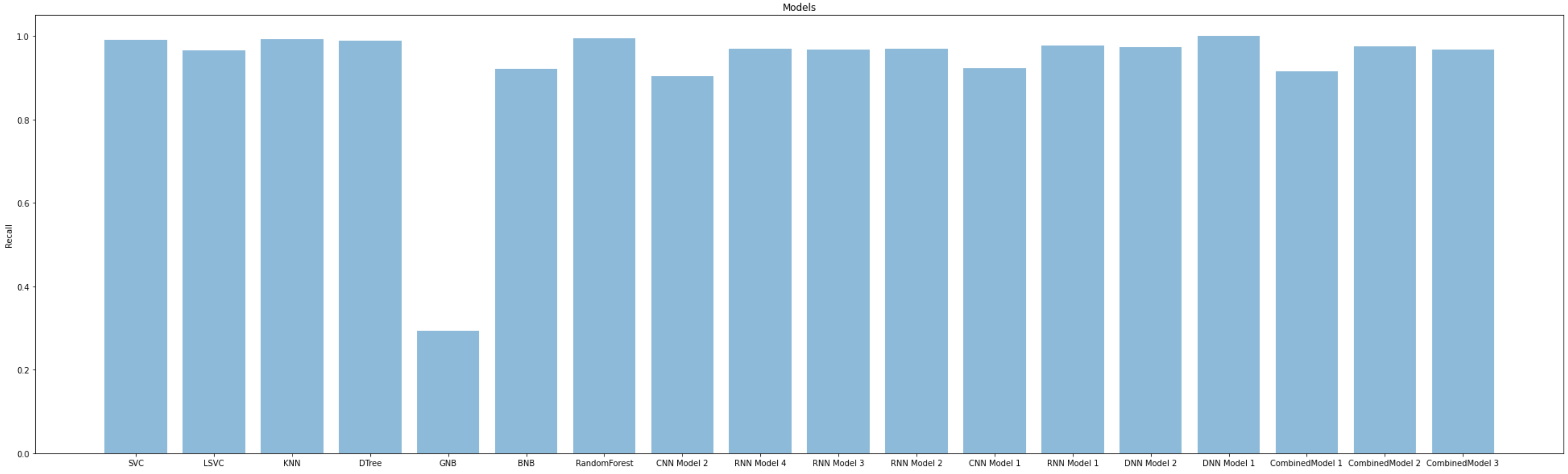
6. EXPERIMENTAL RESULTS – BEST ACCURACY



6. EXPERIMENTAL RESULTS – BEST F1-SCORE



6. EXPERIMENTAL RESULTS – BEST PRECISION



6. EXPERIMENTAL RESULTS – BEST RECALL

7. CONCLUSIONS

The deep learning algorithm that worked best was RNN. Although Random Forest, Decision Tree, SVM, KNN give good results in all comparison methods, we recommended it because Random Forest gave the best result.

REFERENCES

1. Mohammad, R.M., McCluskey, L., & Thabtah, F. (2015, March). Training Dataset. URL: https://archive.ics.uci.edu/ml/index.php

2. Tan, C.L. (2018, March). Phishing Legitimate Full. URL: https://data.mendeley.com/

3. Banik, S. (2021, January). Dataset. URL: https://www.kaggle.com/

4. Adebowale, M. (2019, December). Text Frame Image Features. URL: https://data.mendeley.com/

5. Hannousse, A., & Yahiouche, S. (2020, September). dataset B 05 2020. URL: https://data.mendeley.com/

6. Mamun, M.S.I., Rathore, M.A., Lashkari, A.H., Stakhanova, N., & Ghorbani, A.A. (2016). Phishing. URL: https://www.unb.ca/

7. Yang, P., Zhao, G., & Zeng, P. (2019, January). Phishing Website Detection Based on Multidimensional Features Driven by Deep Learning (pp. 15196-15209). IEEE.

8. Wang, W., Zhang, F., Luo, X., & Zhang, S. (2019, October). PDRCNN: Precise Phishing Detection with Recurrent Convolutional Neural Networks (pp. 1-16). Wiley-Hindawi.

9. Adebowale, M.A., Lwin, K.T., & Hossain, M.A. (2019, August). Deep Learning with Convolutional Neural Network and Long Short-Term Memory for Phishing Detection. In 13th International Conference on SKIMA (pp. 1-8). IEEE, Island of Ukulhas, Maldives.

10. Kumar, J., Santhanavijayan, A., Janet, B., Rajendran, B., & Bindhumadhava, B.S. (2020, January). Phishing Website Classification and Detection Using Machine Learning. In ICCCI (pp. 1-6). IEEE, Coimbatore, India.

11. Opara, C., Wei, B., & Chen, Y. (2020, July). HTMLPhish: Enabling Phishing Web Page Detection by Applying Deep Learning Techniques on HTML Analysis. In IJCNN (pp. 1-8). IEEE, Glasgow, UK.

12. Yang, S. (2020, June). Research on Website Phishing Detection Based on LSTM RNN. In 4th ITNEC (pp. 284-288). IEEE, Chongqing, China.

THANK YOU FOR LISTENING TO US