CSM501: CLASSIFICATION OF MALWARE USING IMAGE REPRESENTATION

Shinit Shetty	119A1076
Anushka Tawte	119A1090
Sanskruti Wathare	119A1097

Project Guide: Dr. Varsha Patil

TABLE OF CONTENTS

01

INTRODUCTION

Discussing the problem to be addressed

02

LITERATURE SURVEY

Brief about previous work on the topic

03

PROPOSED SYSTEM

Steps completed so far

04

CONCLUSION

Methodology and Results

INTRODUCTION

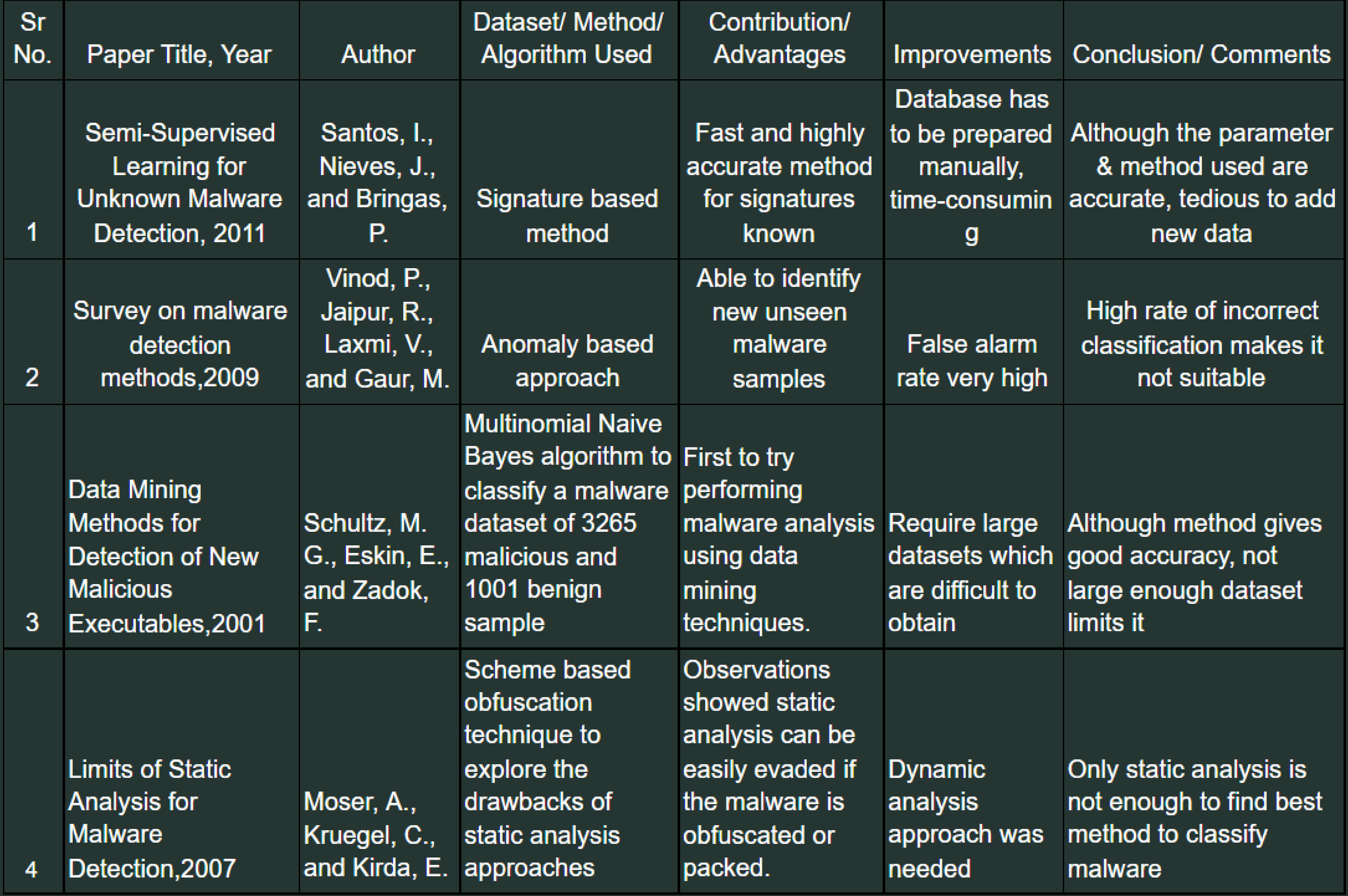
- The Internet: risk of getting attacked.

- Malware detection and classification: one of the most crucial problems in the field of cyber security.

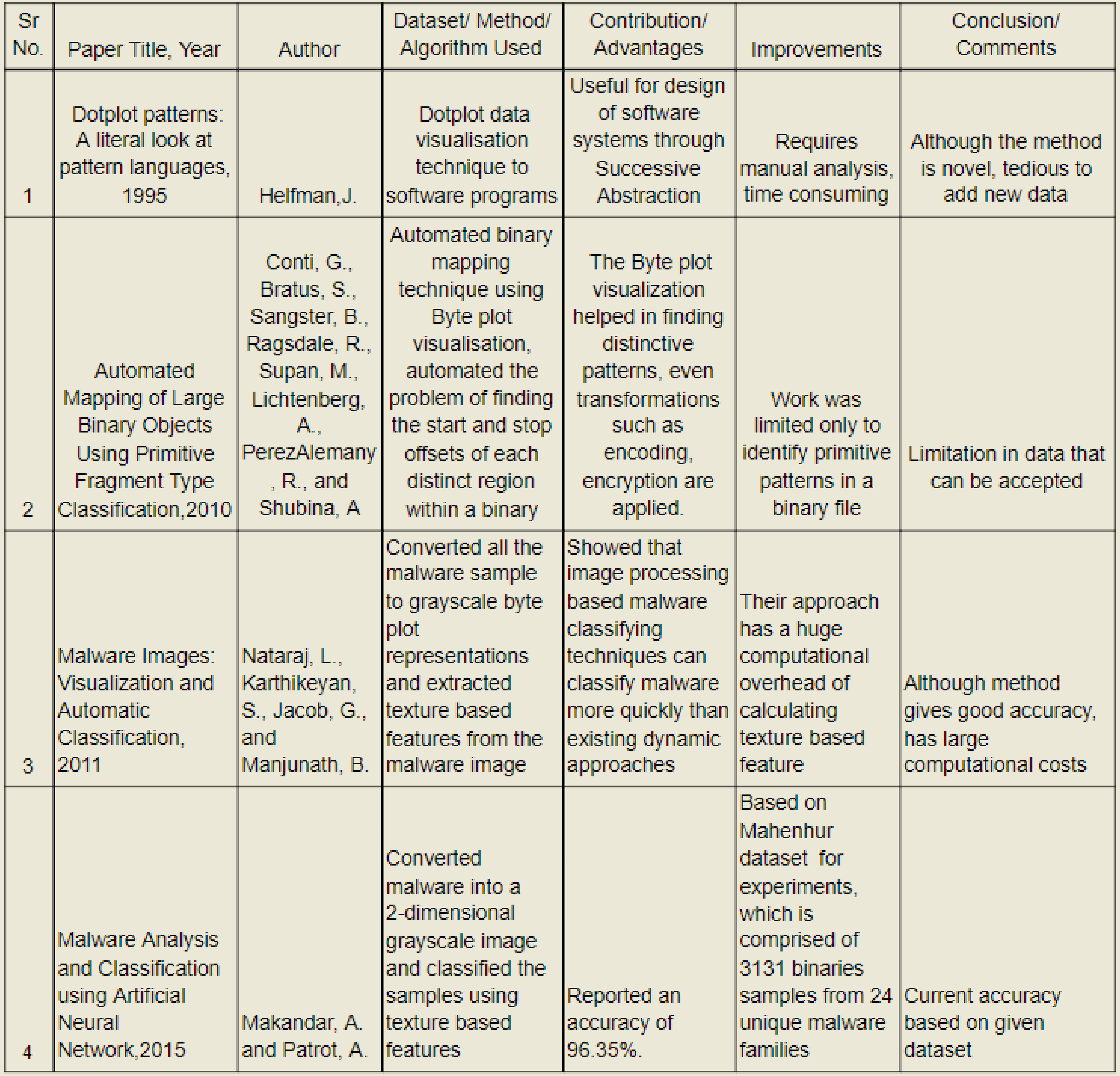
- Malware analysis is the process of dissecting a binary file to understand its working and then devising methods to identify it and other similar files

- Signature-based technique fail to cope with code obfuscation and fail to effectively identify newly arrived threats.

LITERATURE SURVEY



Sr No.	Paper Title, Year	Author	Dataset/ Method/ Algorithm Used	Contribution/ Advantages	Improvements	Conclusion/ Comments
1	Semi-Supervised Learning for Unknown Malware Detection, 2011	Santos, I., Nieves, J., and Bringas, P.	Signature based method	Fast and highly accurate method for signatures known	Database has to be prepared manually, time-consumin g	Although the parameter & method used are accurate, tedious to add new data
2	Survey on malware detection methods,2009	Vinod, P., Jaipur, R., Laxmi, V., and Gaur, M.	Anomaly based approach	Able to identify new unseen malware samples	False alarm rate very high	High rate of incorrect classification makes it not suitable
3	Data Mining Methods for Detection of New Malicious Executables,2001	Schultz, M. G., Eskin, E., and Zadok, F.	Multinomial Naive Bayes algorithm to classify a malware dataset of 3265 malicious and 1001 benign sample	First to try performing malware analysis using data mining techniques.	Require large datasets which are difficult to obtain	Although method gives good accuracy, not large enough dataset limits it
4	Limits of Static Analysis for Malware Detection,2007	Moser, A., Kruegel, C., and Kirda, E.	Scheme based obfuscation technique to explore the drawbacks of static analysis approaches	Observations showed static analysis can be easily evaded if the malware is obfuscated or packed.	Dynamic analysis approach was needed	Only static analysis is not enough to find best method to classify malware



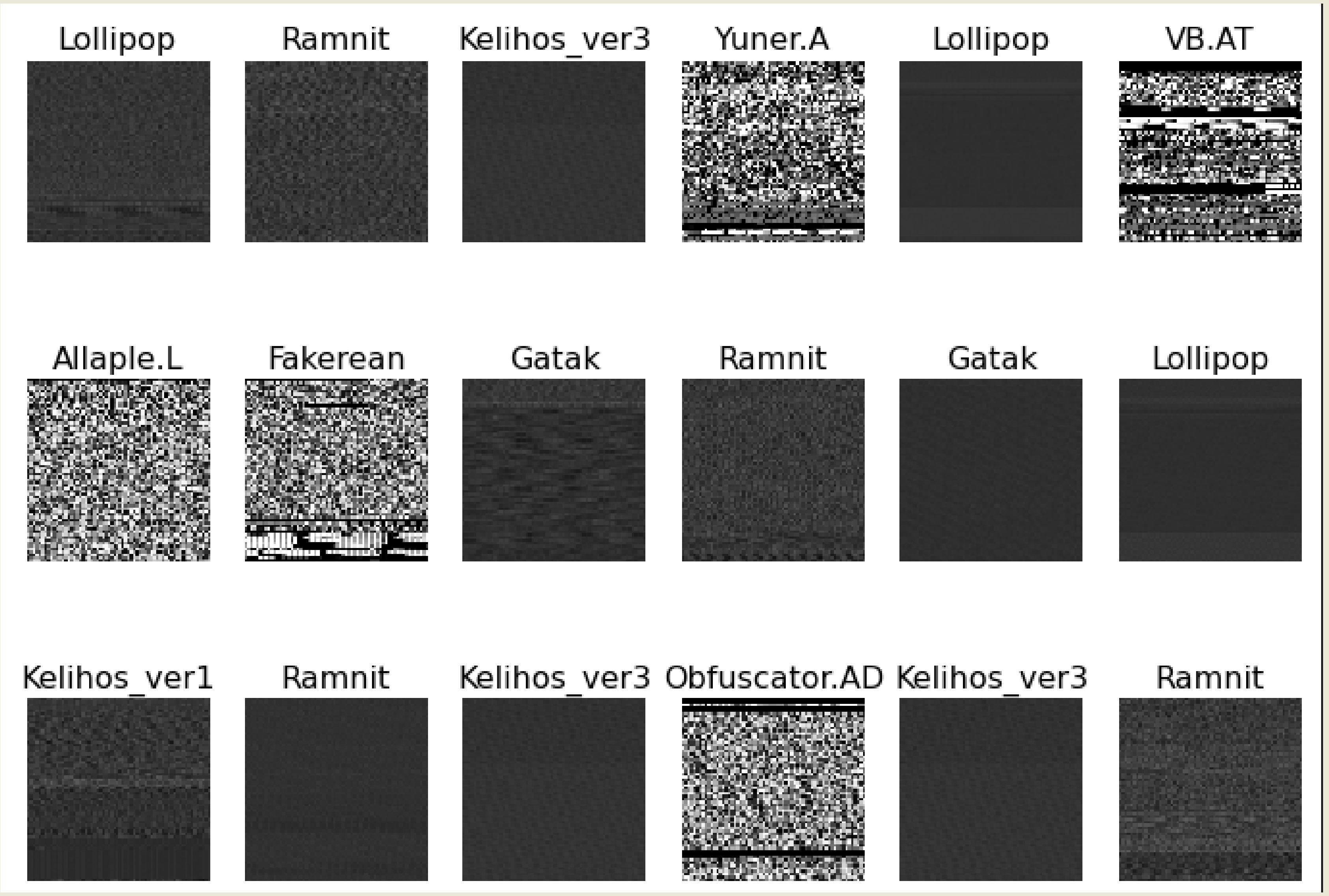
Sr No.	Paper Title, Year	Author	Dataset/ Method/ Algorithm Used	Contribution/ Advantages	Improvements	Conclusion/ Comments
1	Dotplot patterns: A literal look at pattern languages, 1995	Helfman,J.	Dotplot data visualisation technique to software programs	Useful for design of software systems through Successive Abstraction	Requires manual analysis, time consuming	Although the method is novel, tedious to add new data
2	Automated Mapping of Large Binary Objects Using Primitive Fragment Type Classification,2010	Conti, G., Bratus, S., Sangster, B., Ragsdale, R., Supan, M., Lichtenberg, A., PerezAlemany, R., and Shubina, A	Automated binary mapping technique using Byte plot visualisation, automated the problem of finding the start and stop offsets of each distinct region within a binary	The Byte plot visualization helped in finding distinctive patterns, even transformations such as encoding, encryption are applied.	Work was limited only to identify primitive patterns in a binary file	Limitation in data that can be accepted
3	Malware Images: Visualization and Automatic Classification, 2011	Nataraj, L., Karthikeyan, S., Jacob, G., and Manjunath, B.	Converted all the malware sample to grayscale byte plot representations and extracted texture based features from the malware image	Showed that image processing based malware classifying techniques can classify malware more quickly than existing dynamic approaches	Their approach has a huge computational overhead of calculating texture based feature	Although method gives good accuracy, has large computational costs
4	Malware Analysis and Classification using Artificial Neural Network,2015	Makandar, A. and Patrot, A.	Converted malware into a 2-dimensional grayscale image and classified the samples using texture based features	Reported an accuracy of 96.35%.	Based on Mahenhur dataset for experiments, which is comprised of 3131 binaries samples from 24 unique malware families	Current accuracy based on given dataset

PROPOSED SYSTEM

PROPOSED SYSTEM

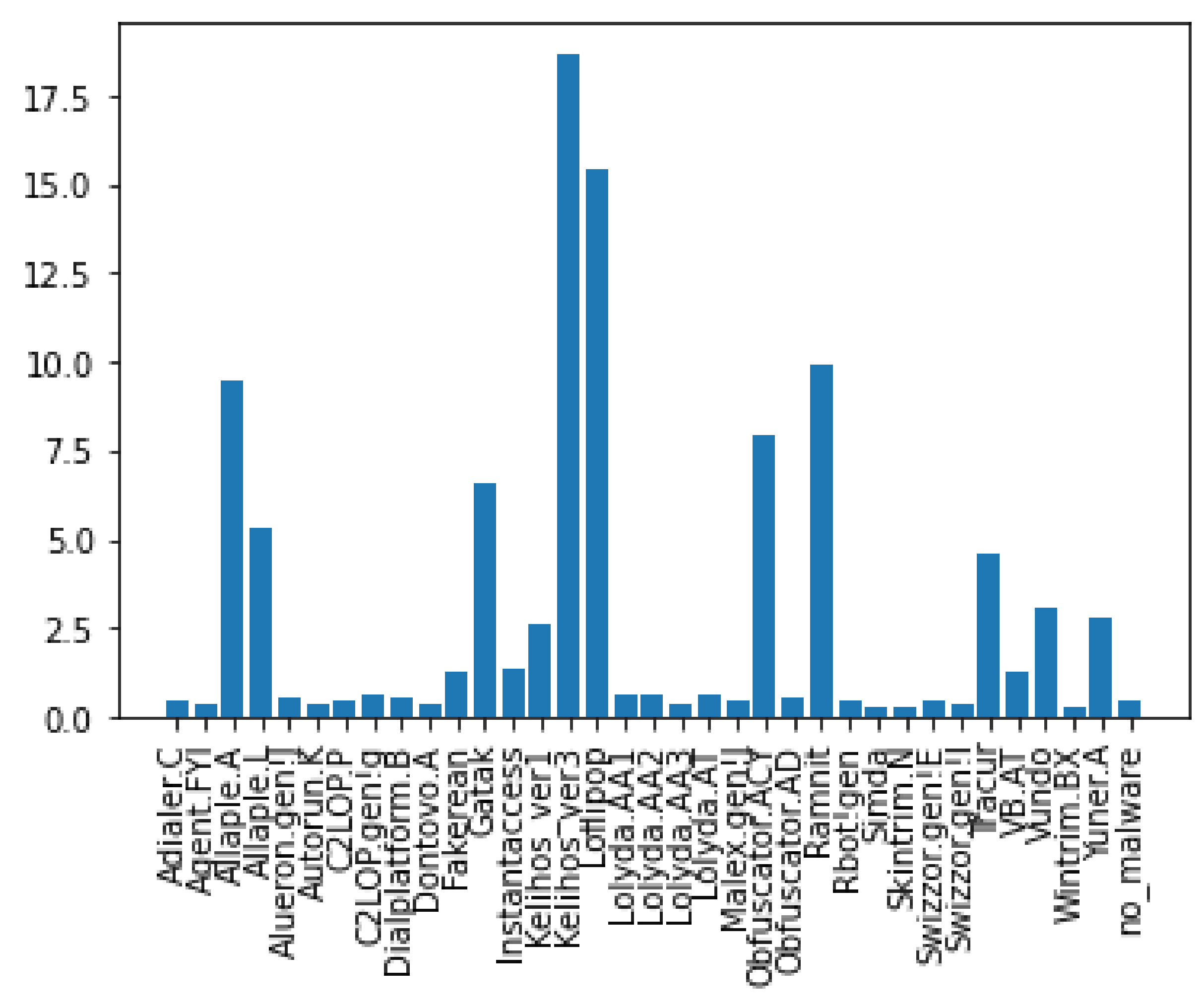
- A novel approach is to use image representation of binaries for classification.

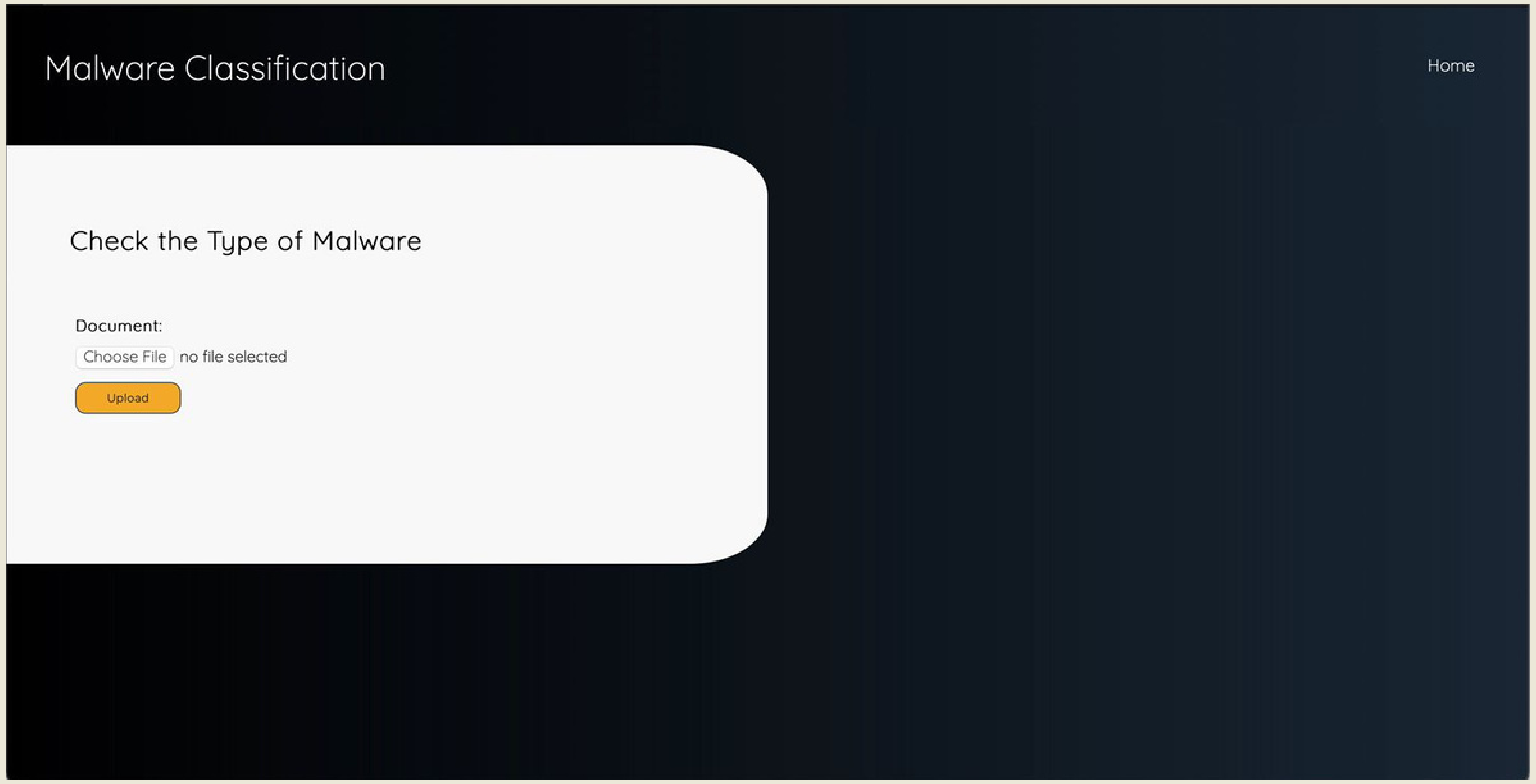
- Machine learning algorithms can be used to classify the files on the basis of features of image



- 34 malware families + non malware

- A comprehensive solution is a website that allows for easy classification of malware




Malware Classification

Check the Type of Malware

Document:

METHODOLOGY

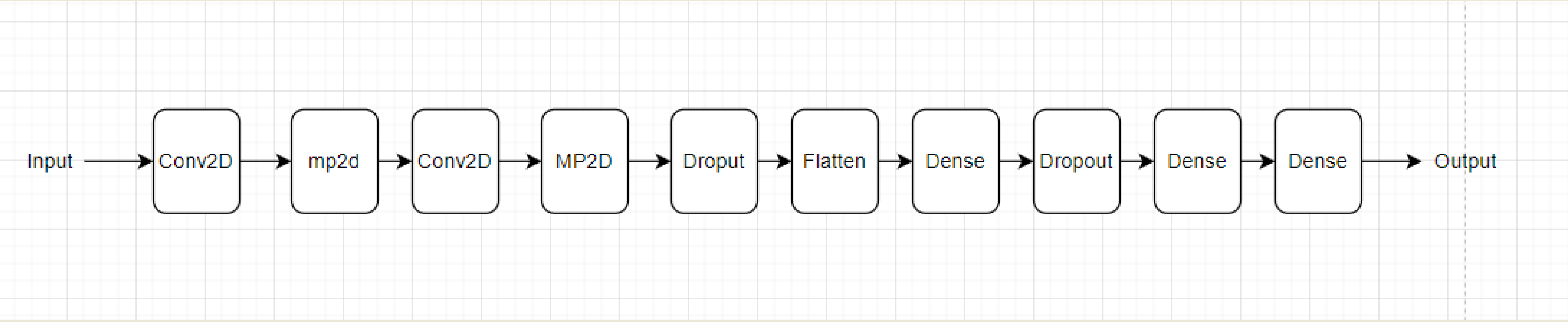
Epochs: 70

Kernel Size: 3 x 3

Accuracy: 86%

Activation Function: 3 Relu, 1 Softmax

No. of layers: 10



OUR CONTRIBUTION

- Added images for non malware files
- Added new malware families for it to adapt to the ever increasing newly arrived threats
- New dataset comprising of malware and non malware files to be uploaded on Kaggle

THANK YOU