A New Learning Approach to Malware Classification Using Discriminative Feature Extraction

Due to internet lots of malware are spread into personal and business computers and they can corrupt files or steal information from infected computer and to prevent this all existing techniques were using local or global features from malware images and if any malware changes happen in local or global features then prediction/classification accuracy will be affected and to avoid this problem author is using multilayer techniques with local and global features to perform accurate prediction in case of changes on local or global features.

In propose paper malware dataset will be converted to binary images and then this images features will be processed through 4 layers for accurate prediction and in propose paper author is comparing propose work called Multilayer Dense SIFT and Multilayer LBP with GIST, SIFT and LBP.

1. LBP is the existing technique which extract only local features from images and due to that reason its prediction accuracy will be less.
2. GIST or SIFT is also existing technique which trained machine learning algorithms by extracting global features.
3. Propose Multilayer LBP or SIFT works in below four layers

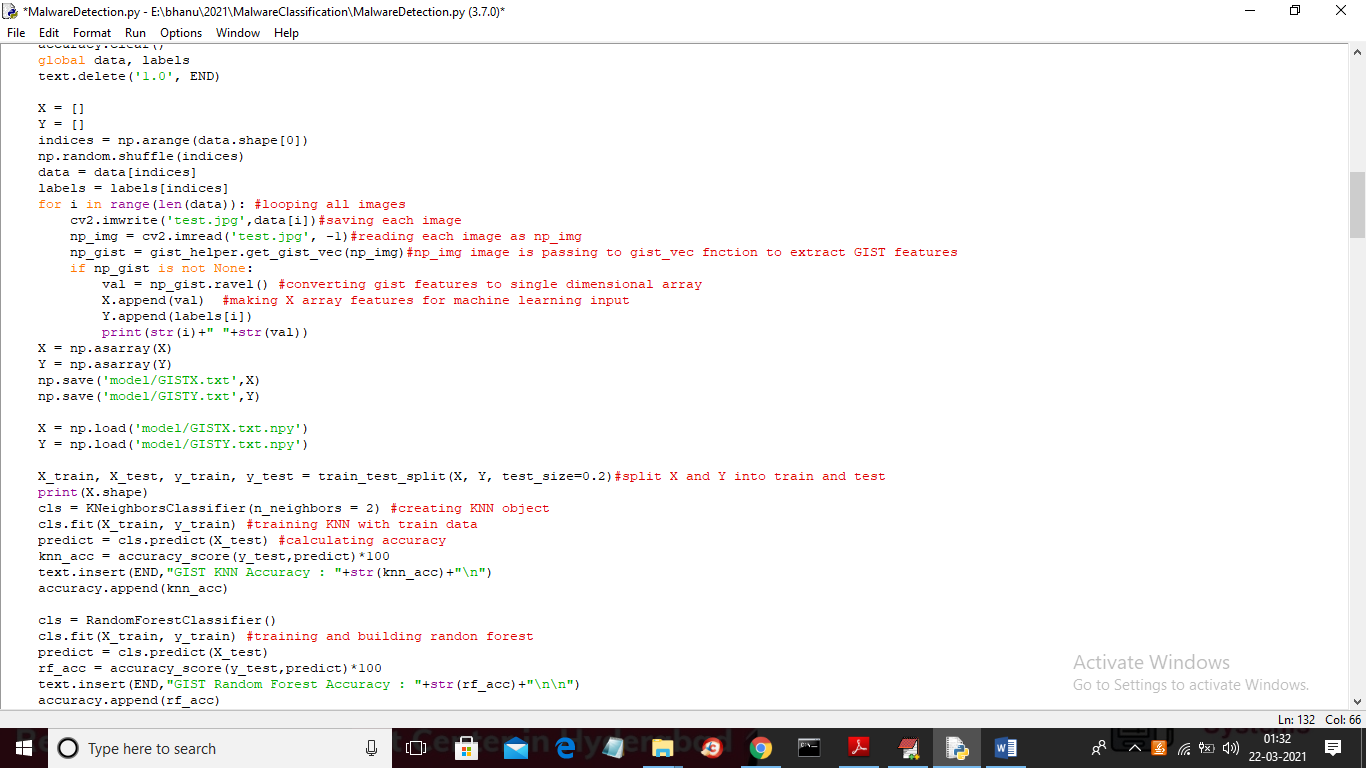
Layer 1): in this layer we will apply either LBP or SIFT to extract features from malware images

Layer 2): To have accurate features image will be split into multiple blocks and then gather relevant features

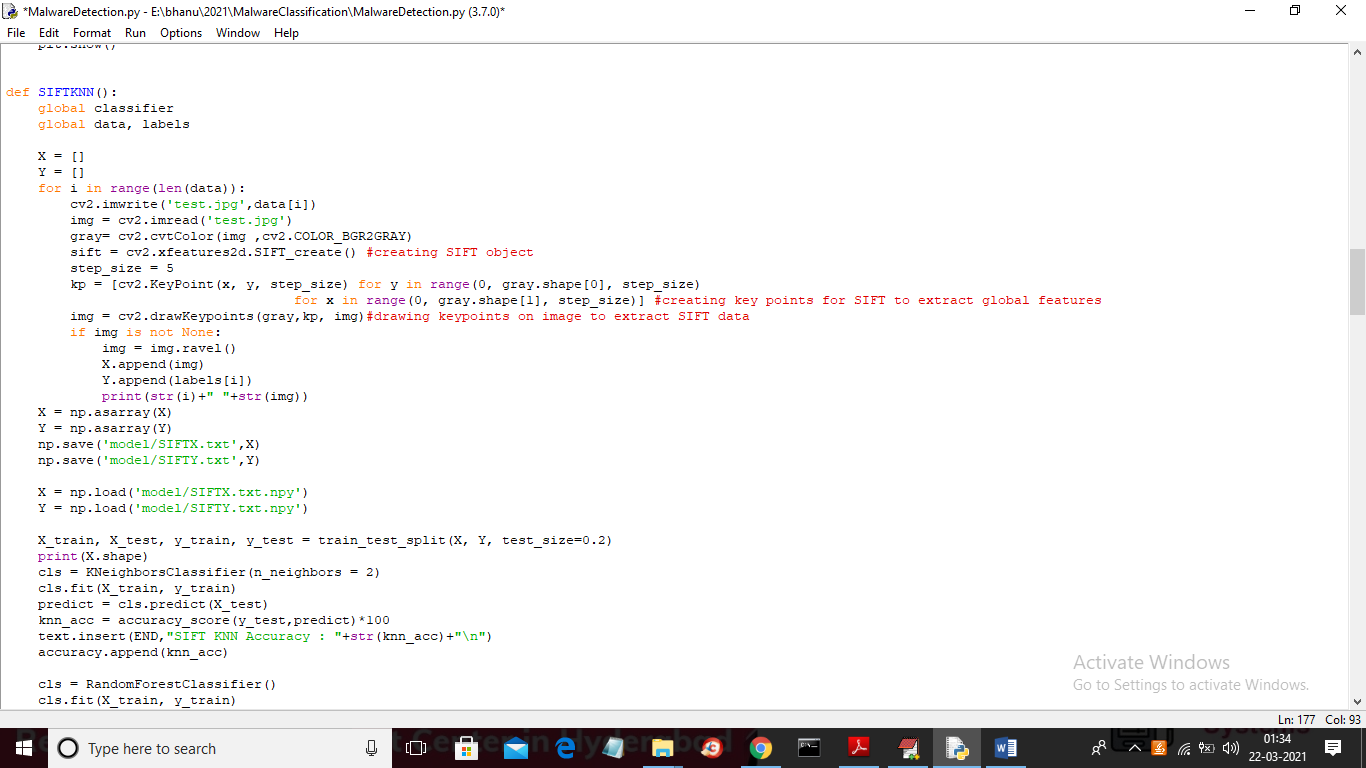
Layer 3): KMEANS algorithm will be applied on accurate features to form clusters

Layer 4) gather or extract all important features from KMEANS and form a Bag Of WORDS vector and this vector will be feed to either KNN or Random Forest to calculate its prediction accuracy.

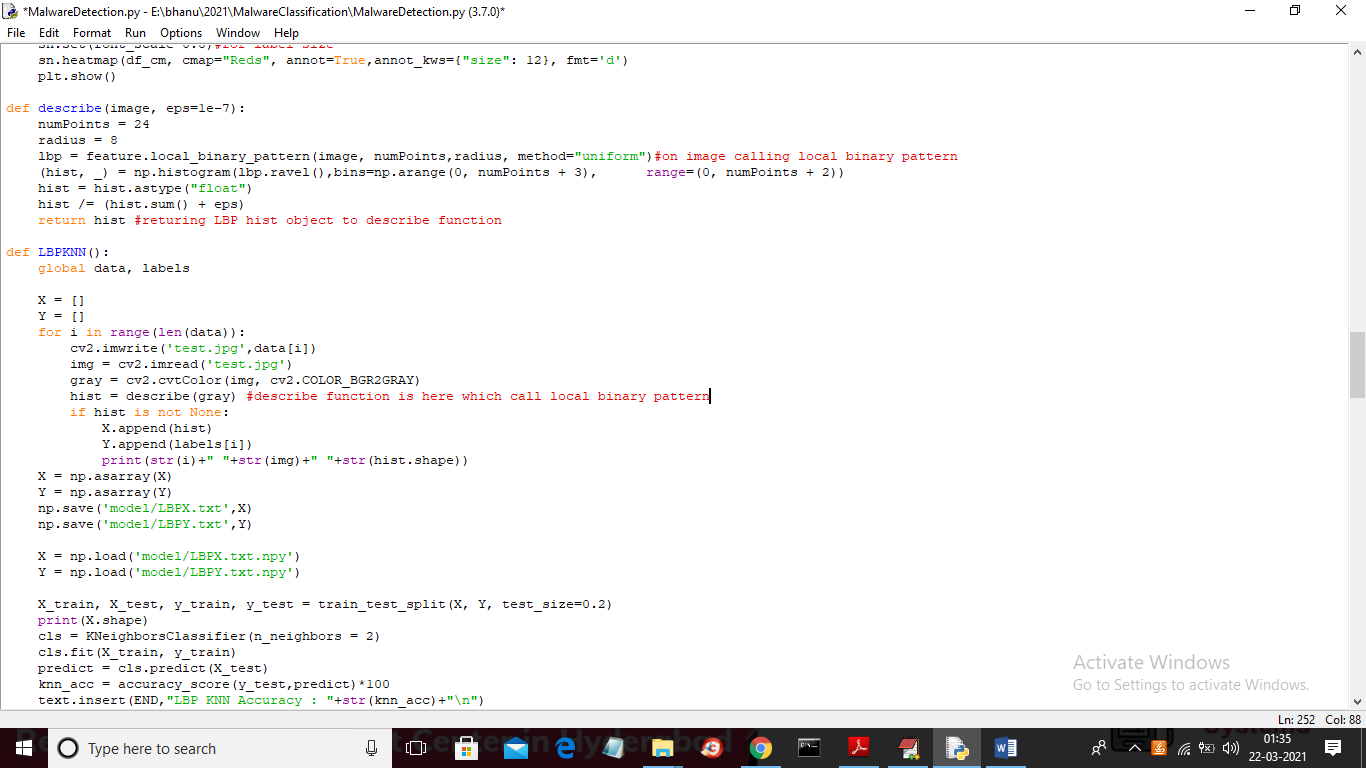
To implement this project we are using MALIMG malware dataset as the dataset describe by author is not available on internet. MALIMG contains nearly 10000 records and each records contains 1024 features from 25 different malware families. Normal laptops will take hours of time to extract GIST, SIFT and LBP features from such huge dataset. So to reduce processing time I extracted 350 malware images from dataset belongs to 3 different malware families. This 350 images can be run in short time. Below code showing how we are extract LBP, SIFT, and GIST features



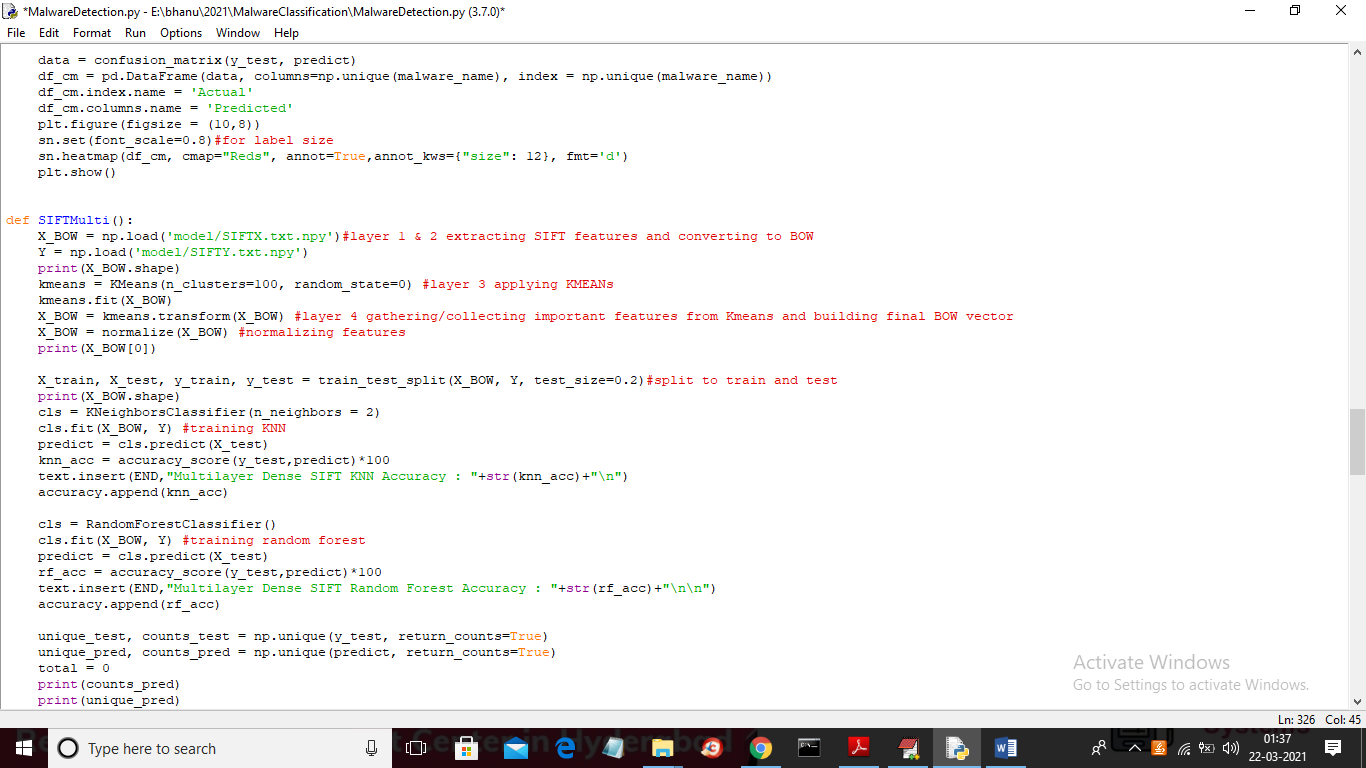
In above screen in comments u can read about GIST features extraction and training KNN and random forest. In below screen u can see about SIFT



In above screen read red colour comments and in below screen you can see about LBP features



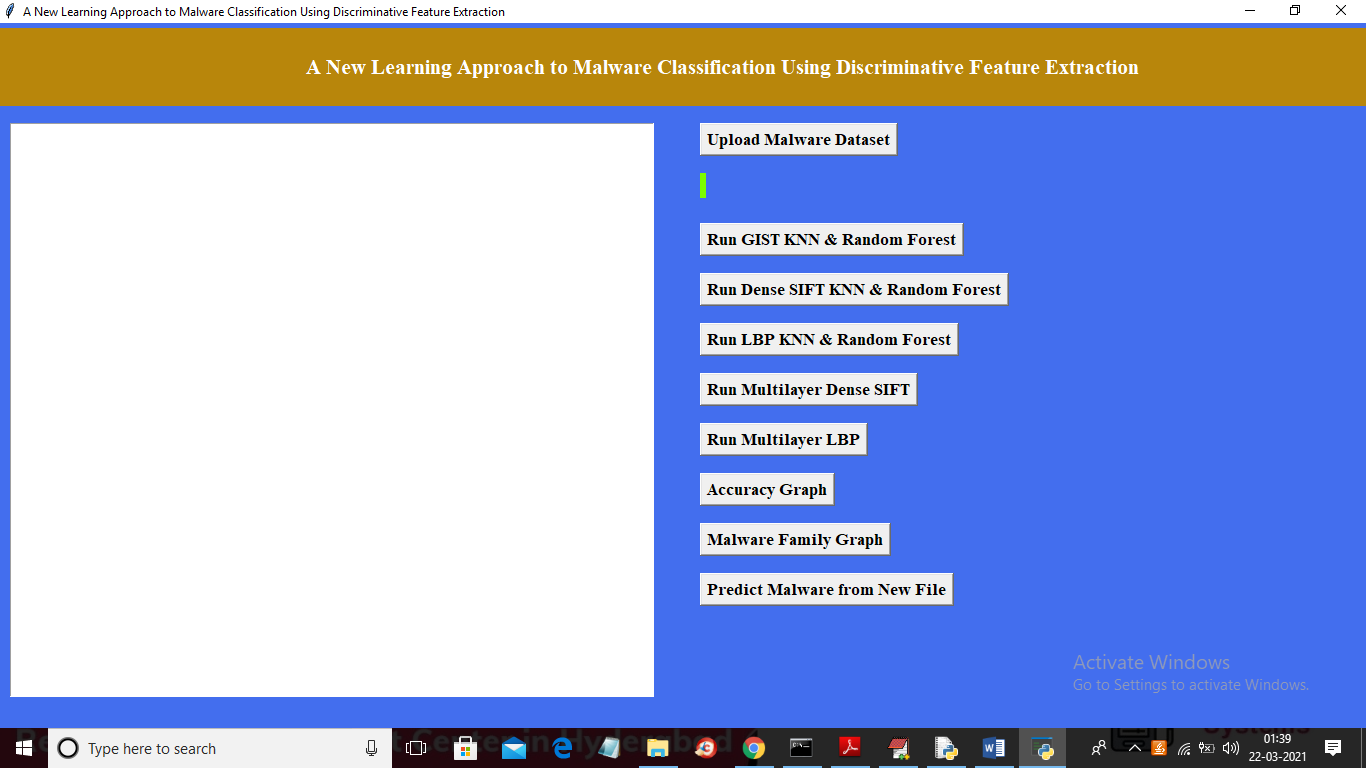
In above screen read comments to see about LBP features and in below screen u can see multilayer propose SIFT dense code



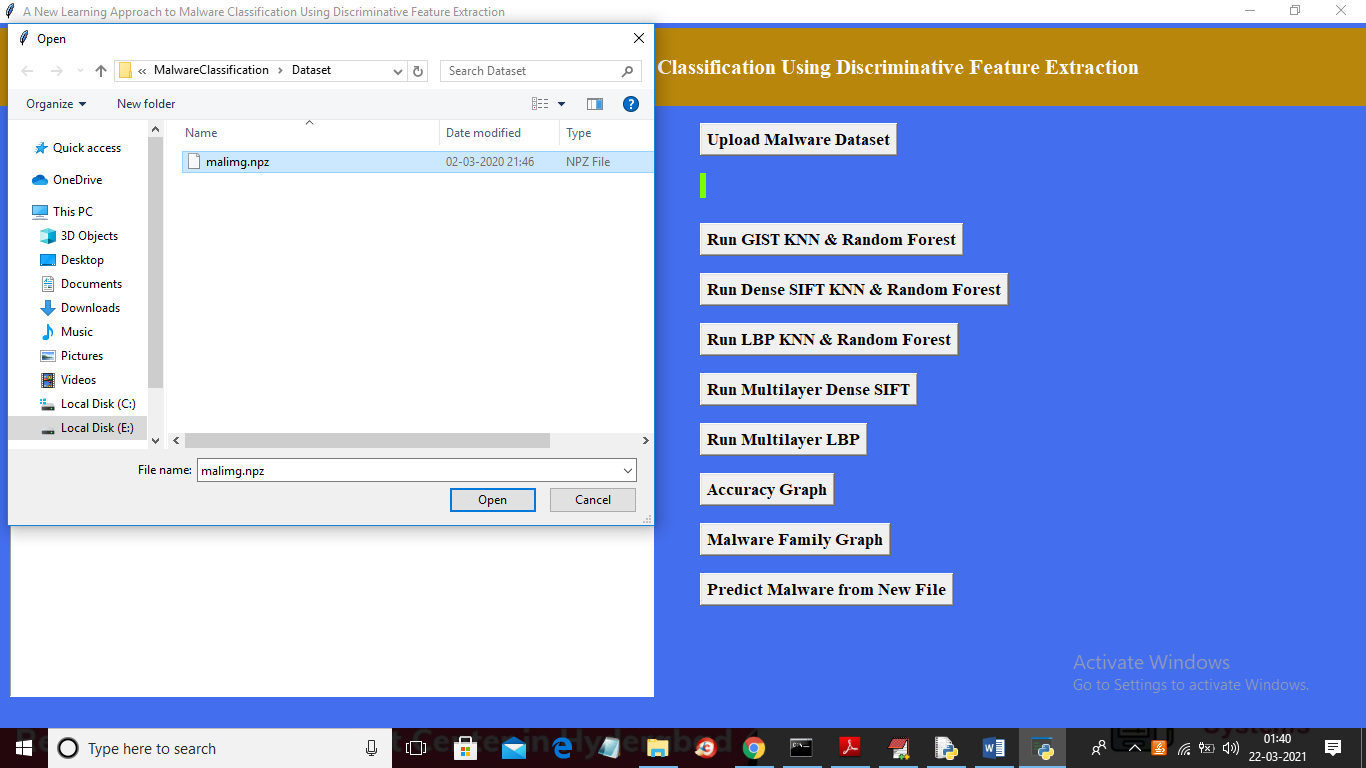
In above screen you can see read comments about 4 layers to extract local and global features as bag of words.

SCREEN SHOTS

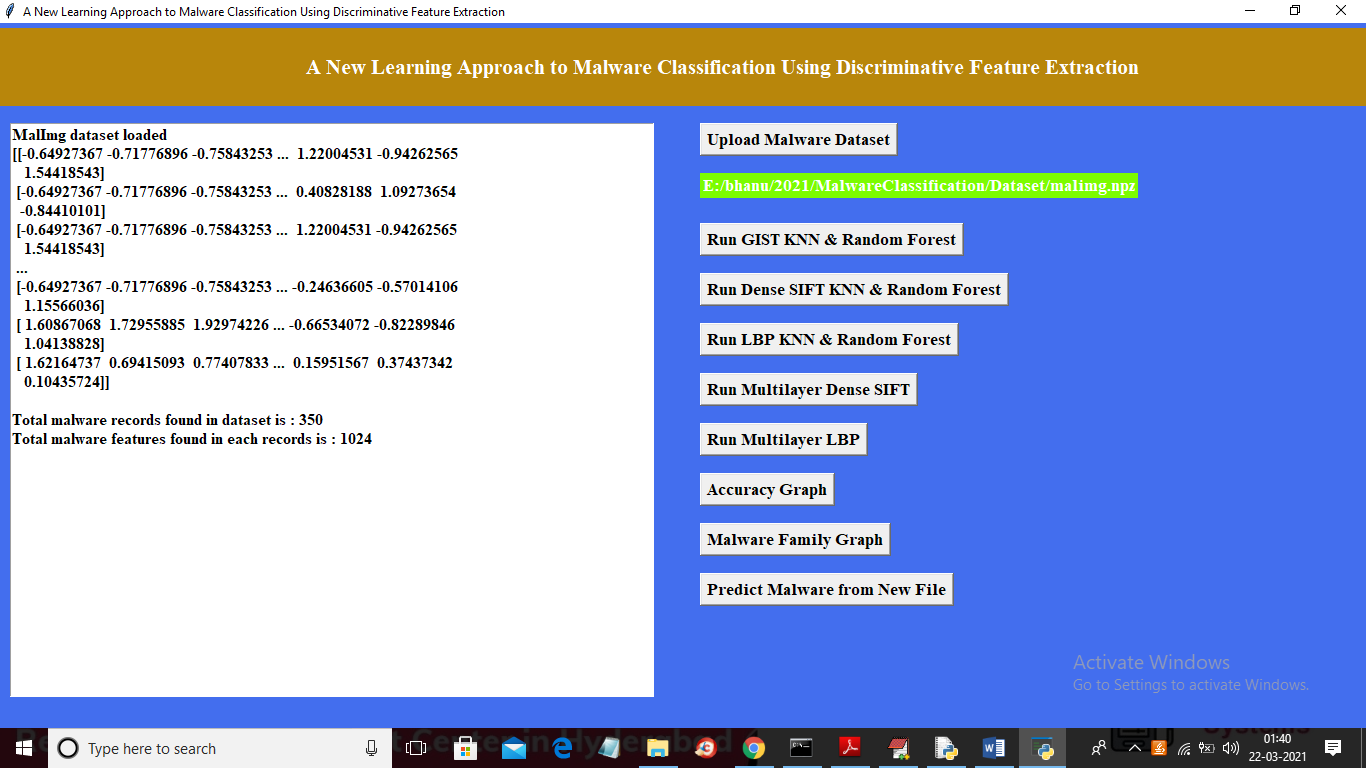
To run project double click on ‘run.bat’ file to get below screen



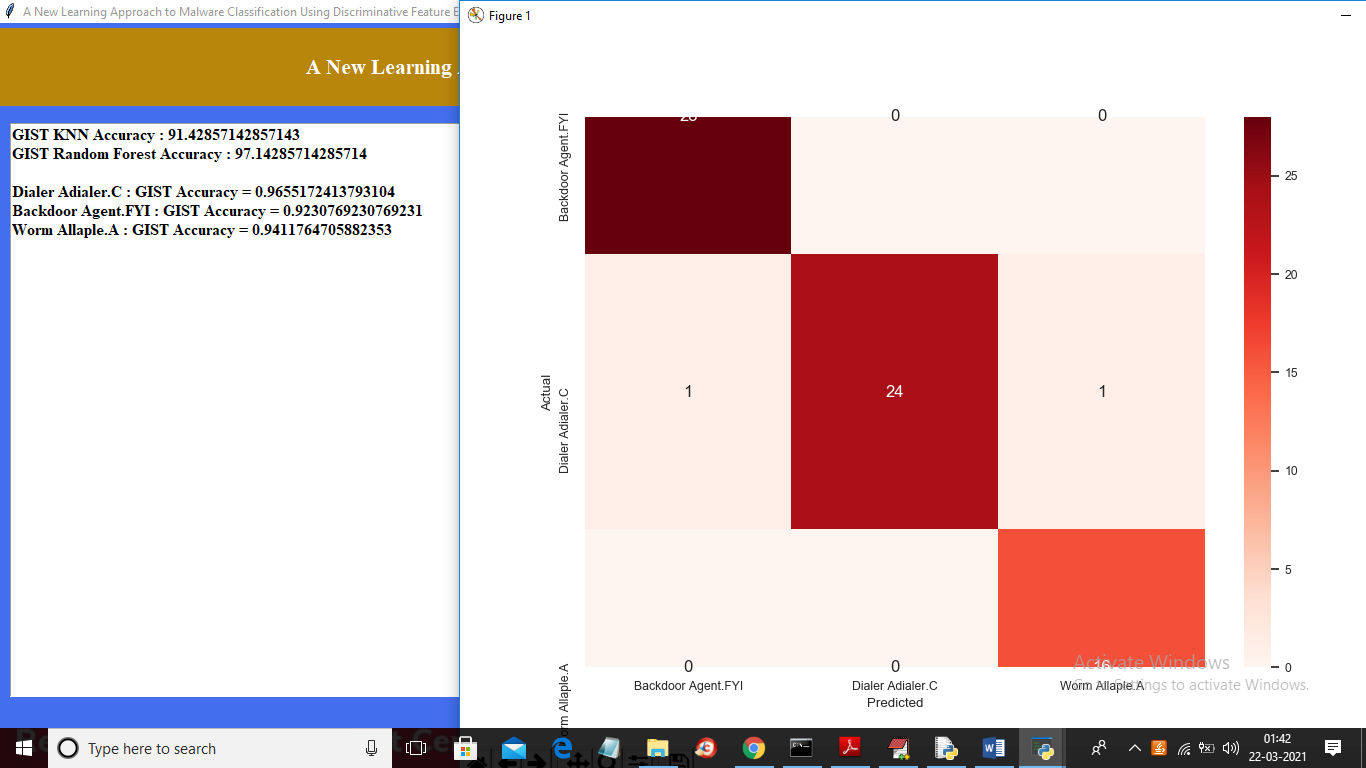
In above screen click on ‘Upload Malware Dataset’ button and upload MALIMG dataset file



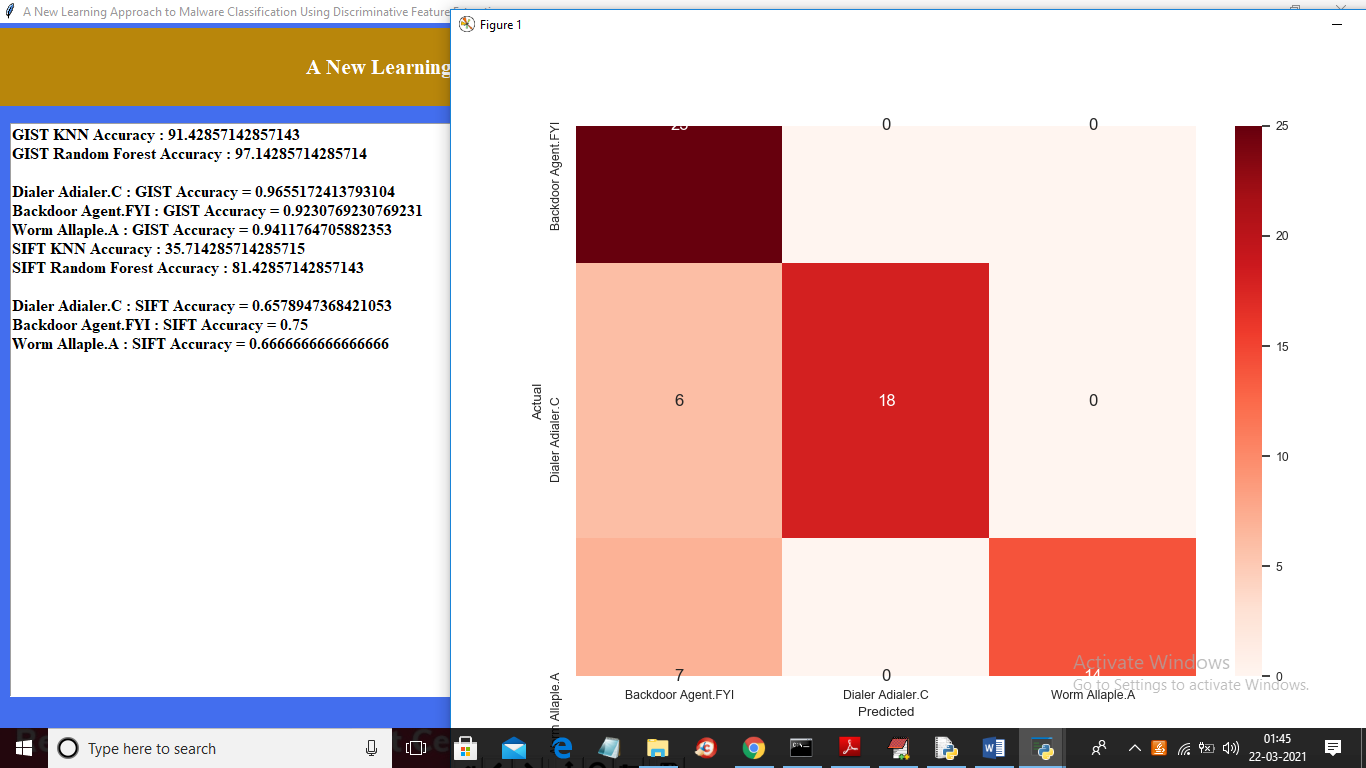
In above screen selecting and uploading malimz.npy file and then click on ‘Open’ button to load dataset and to get below screen



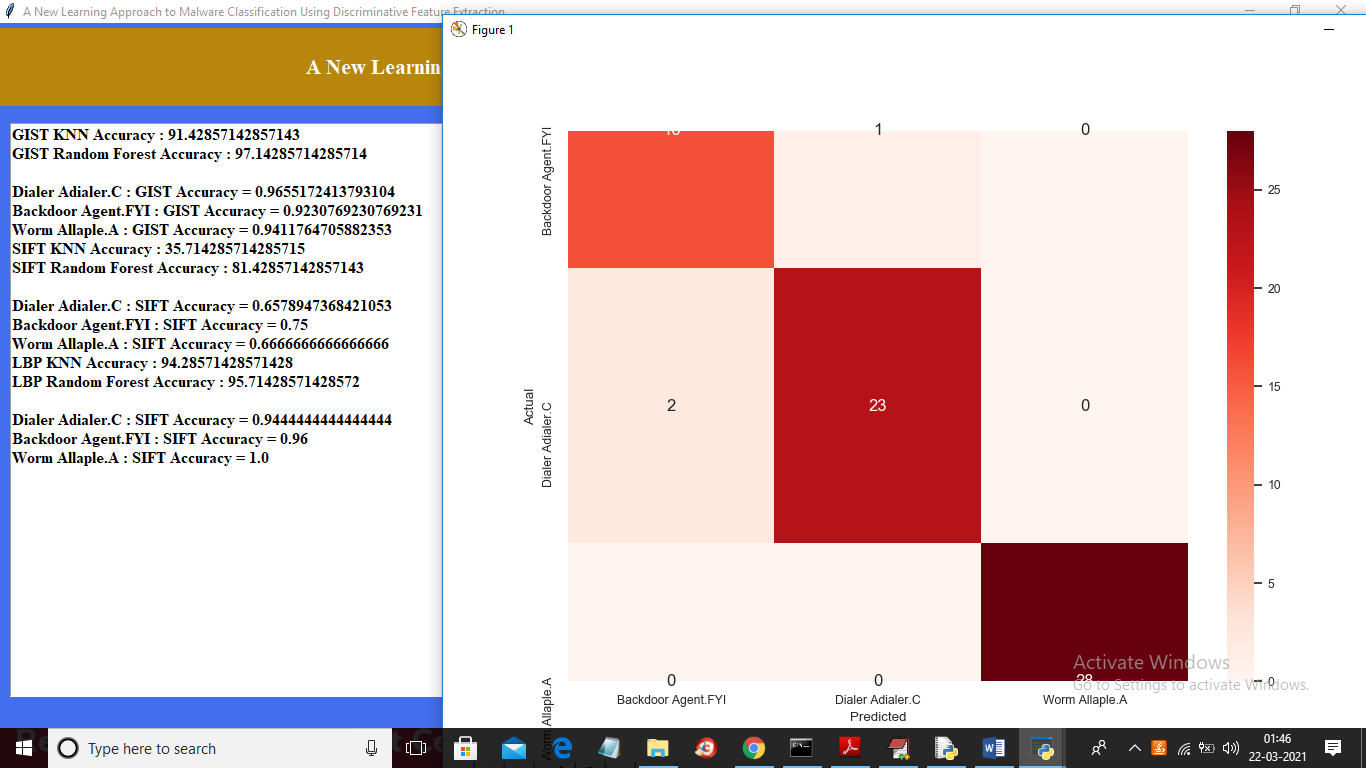
In above screen in extracted 350 malware images and each image contains 1024 features or pixels and now click on ‘run GIST KNN & Random Forest’ button to extract GIST features from all 350 images and then apply KNN and Random Forest algorithm



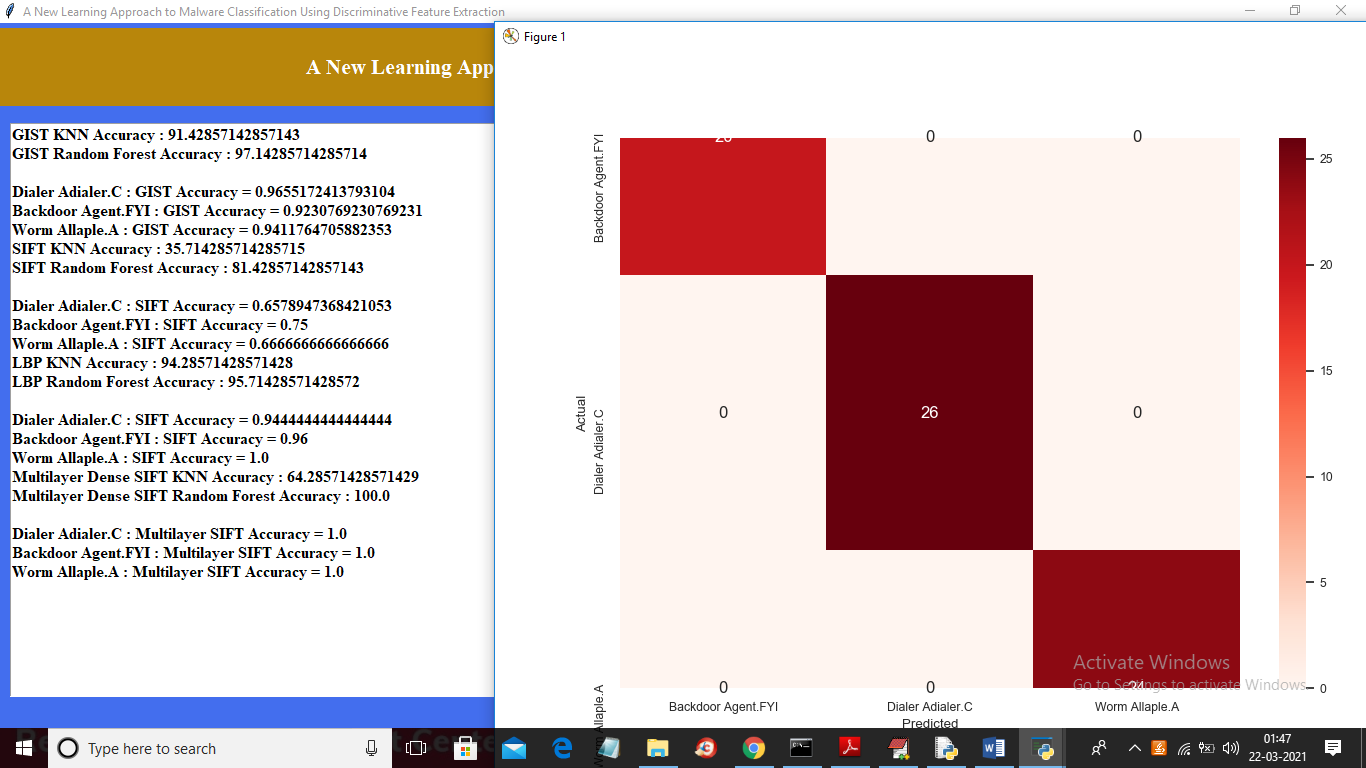
In above screen we can see GIST KNN and random forest accuracy and then we can see accuracy of 3 different malware family and in graph in confusion matrix we can see how many records classified in each malware family for example in 24 records classified as ‘Dialer Adilaer.C’ malware. Now click on ‘Run Dense SIFT KNN & Random Forest’ to extract SIFT features and apply Random forest and KNN



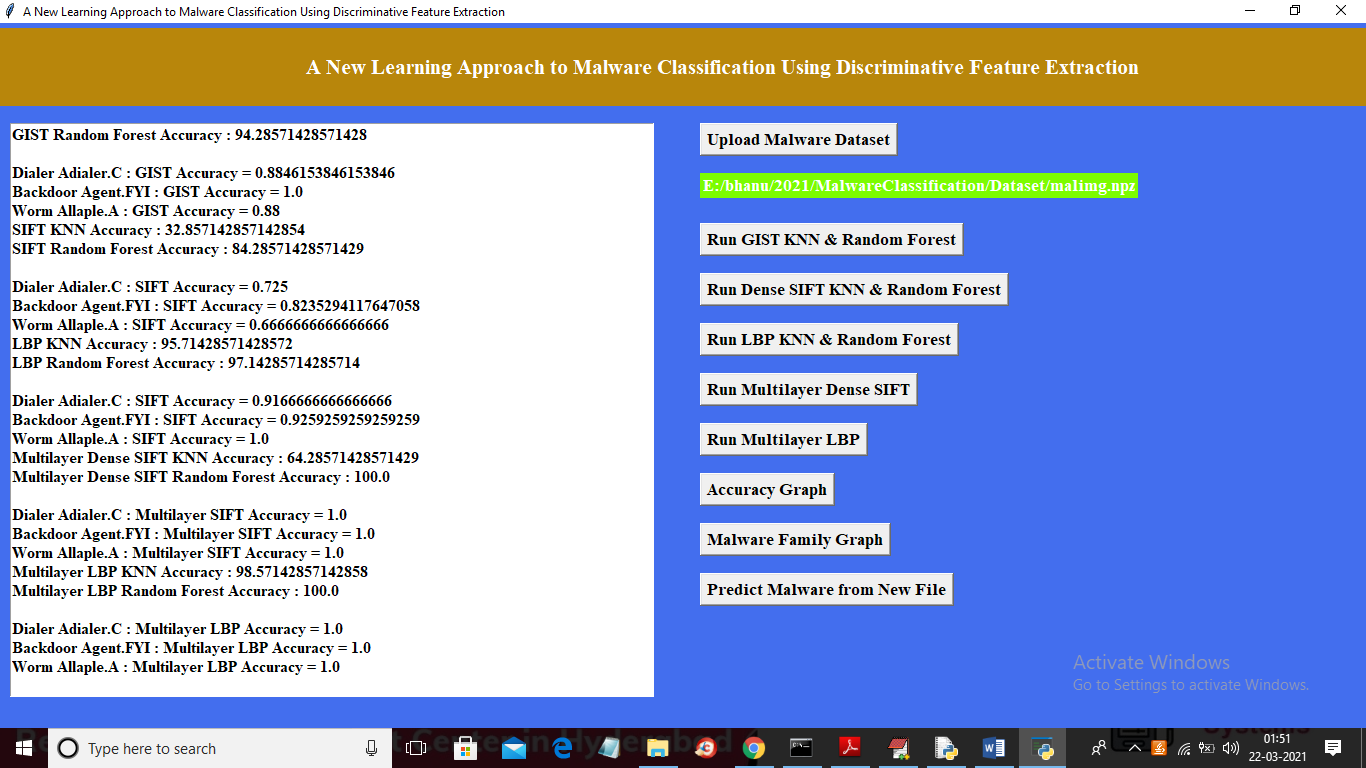
In above screen we can see SIFT KNN and random forest accuracy and now click on ‘Run LBP KNN & Random Forest’ button to extract LBP features and apply both algorithms



In above screen LBK KNN got 94 and random forest got 95% accuracy and now click on ‘Run Multilayer Dense SIFT’ button to apply multilayer SIFT technique and then run both algorithms



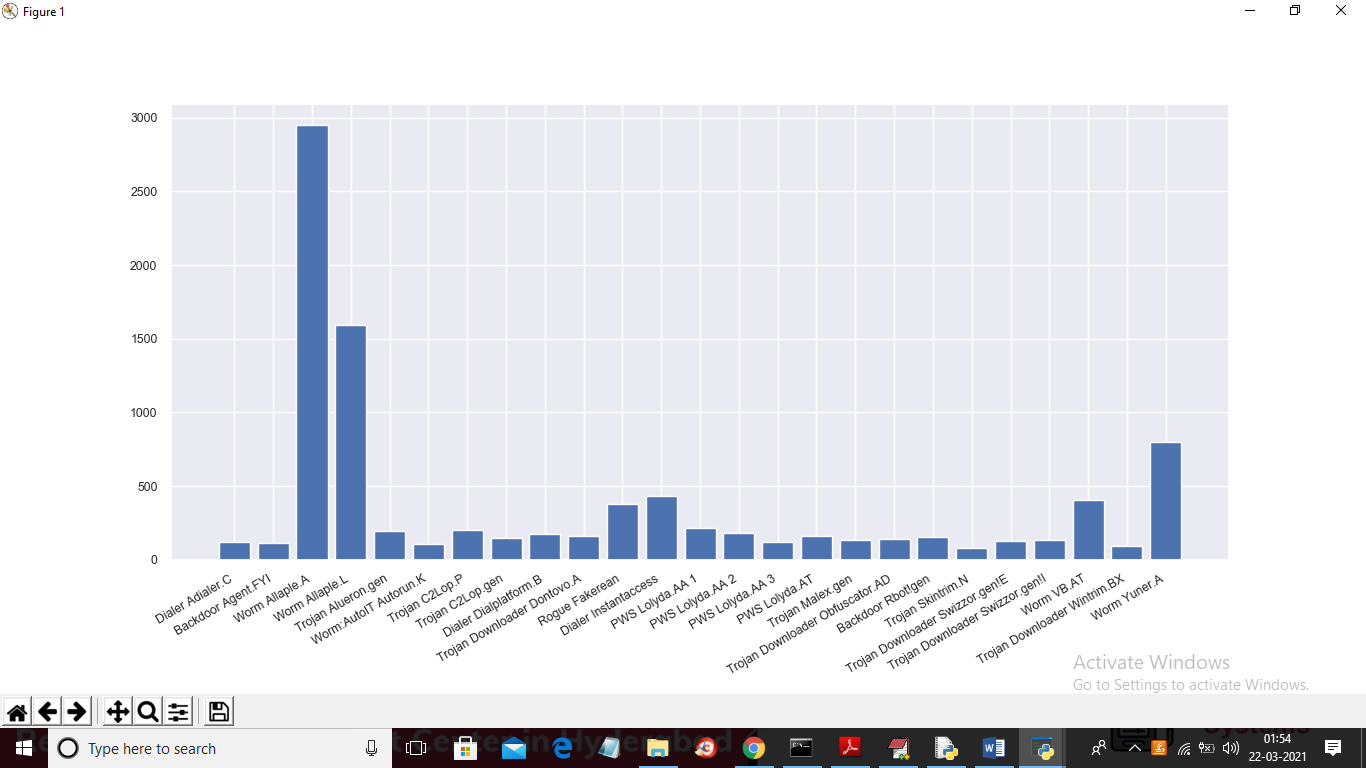
In above screen multilayer SIFT KNN got 64% accuracy and random forest got 100% accuracy and now click on ‘Run Multilayer LBP’ button to apply multilayer technique and then apply both algorithms to get below result



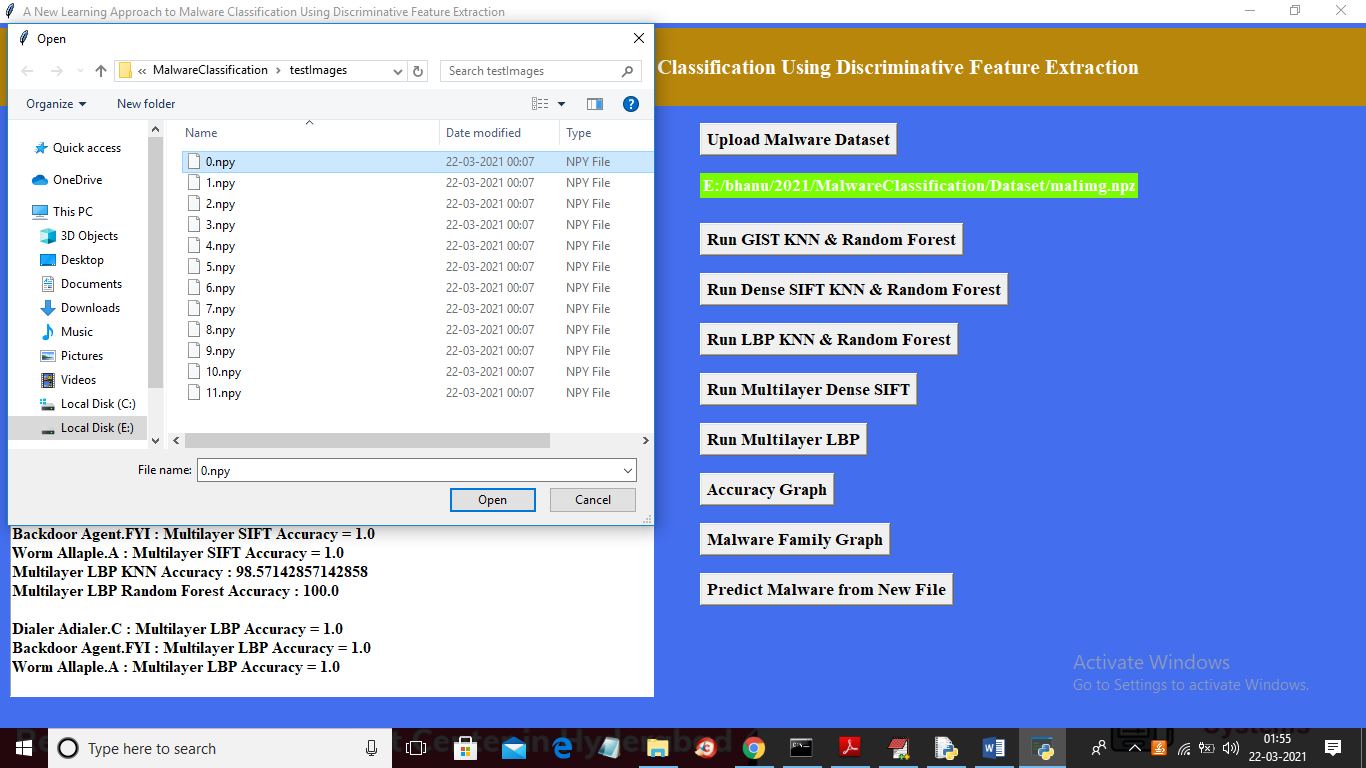
In above screen LBP KNN got 98% accuracy and random forest got 100% accuracy and now click on ‘Accuracy Graph’ button to get below graph



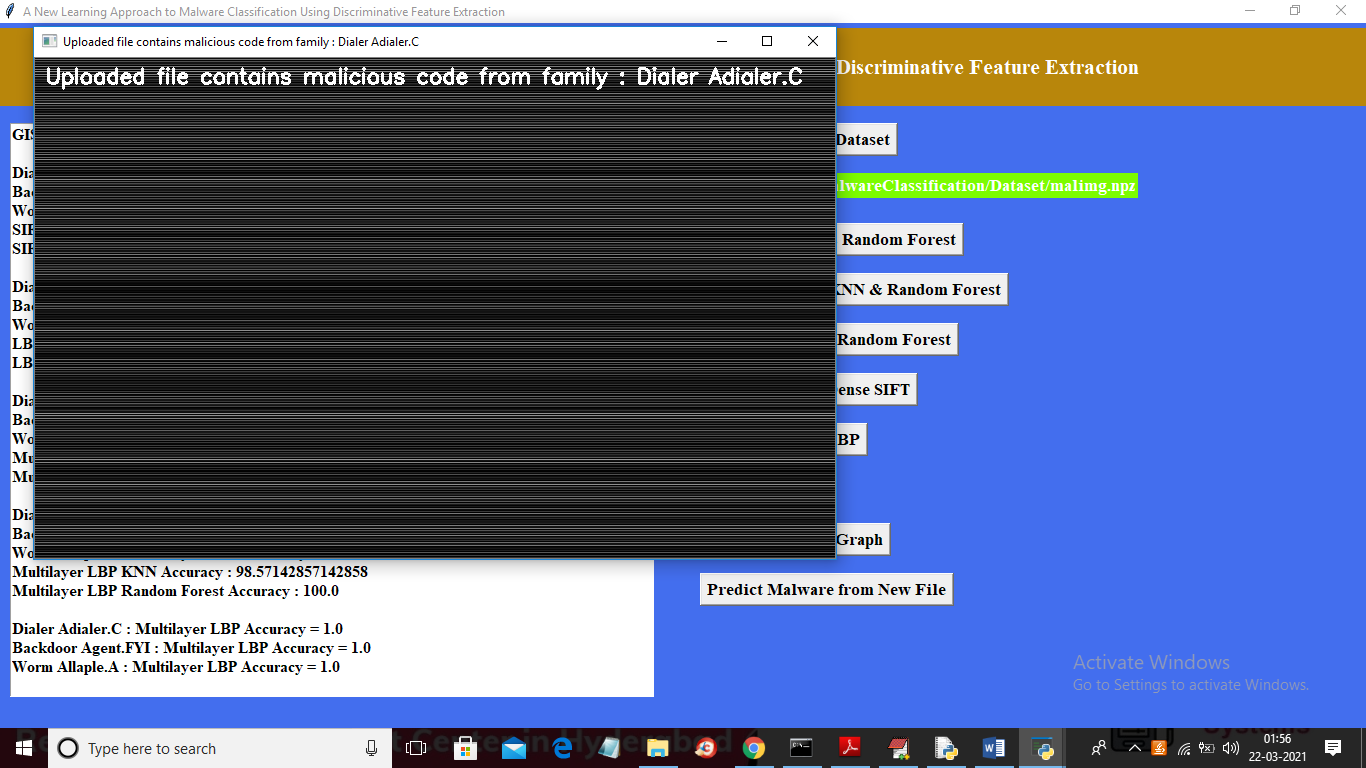
In above graph x-axis represents algorithm name and y-axis represents accuracy and from above graph we can conclude that propose last 3 algorithms perform well compare to existing algorithms and now click on ‘Malware Family Graph’ button to get count of each malware family



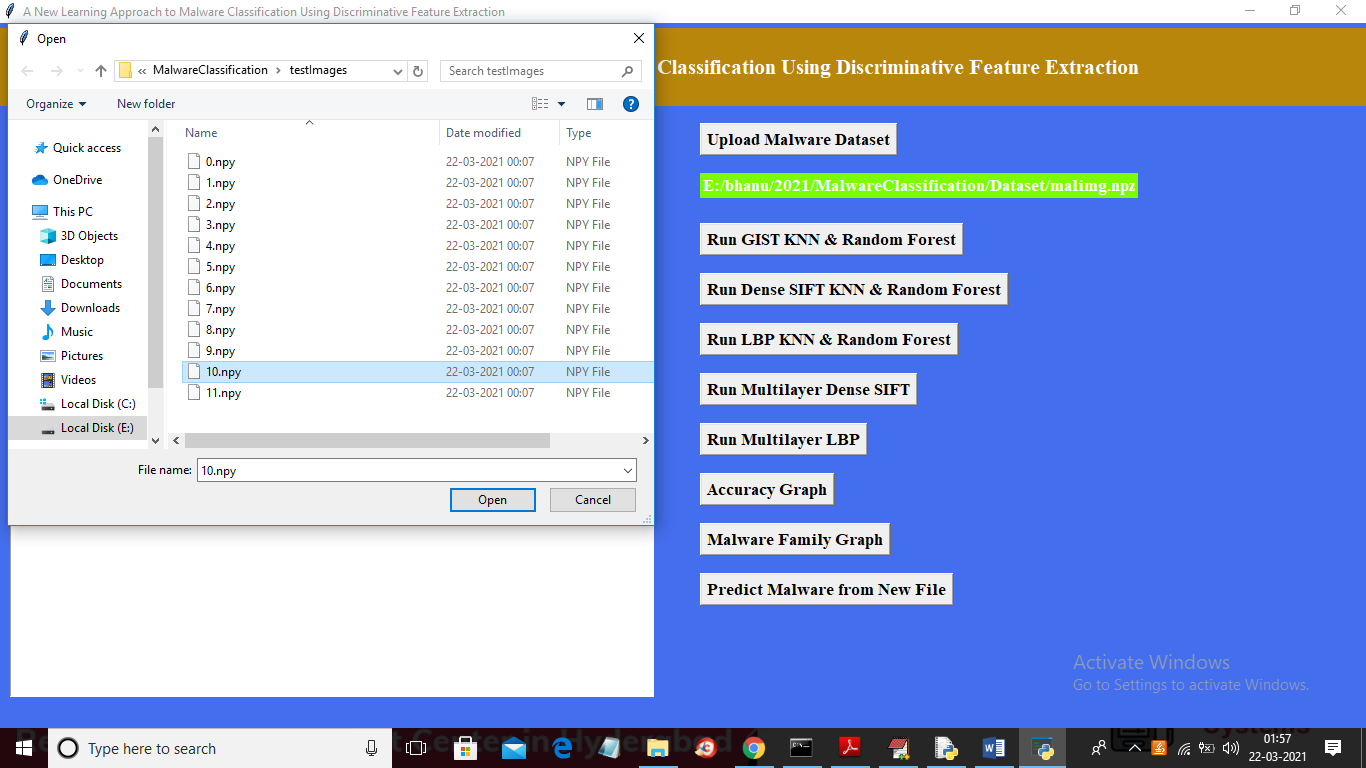
In above graph x-axis represents malware name and y-axis represents count of each malware family and now click on ‘Predict Malware from New File’ button to upload any malware file and then application will predict/classify malware name from that file



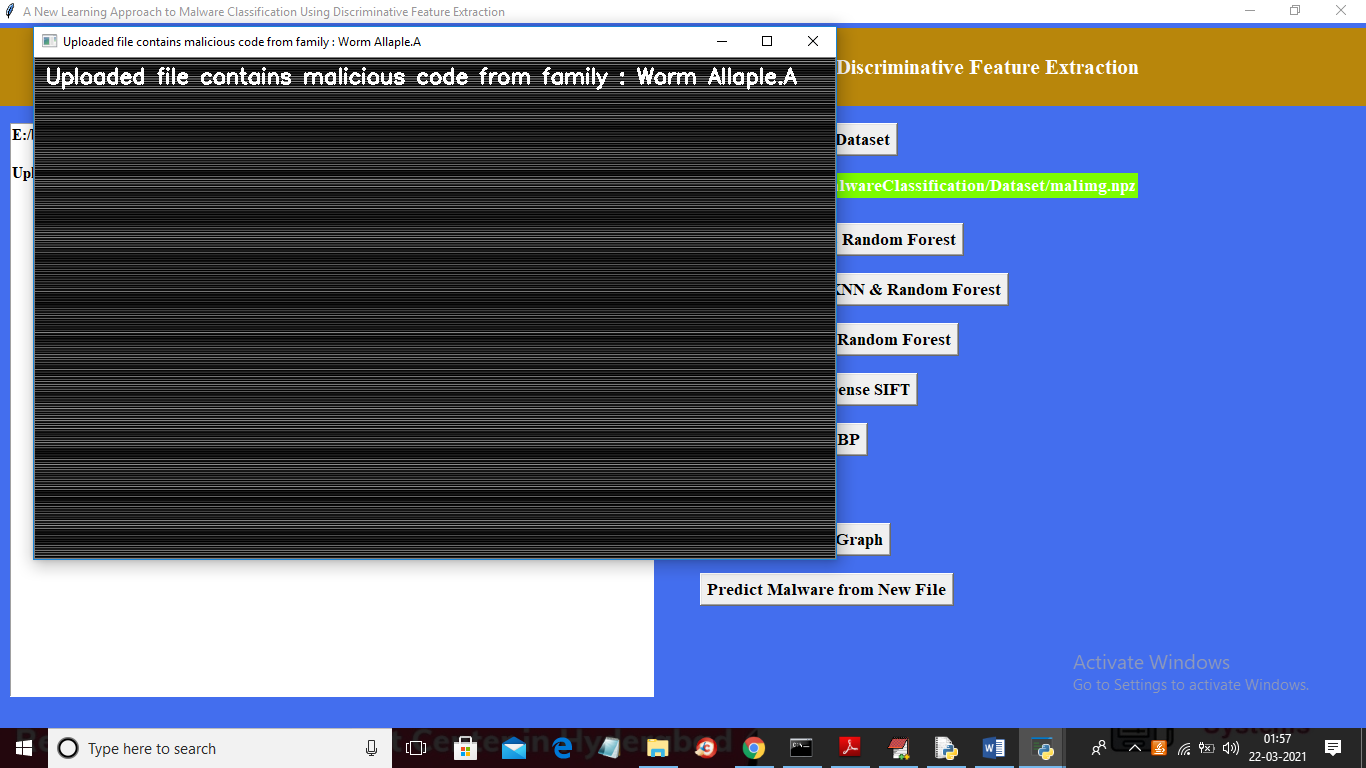
In above screen I am uploading ‘0.npy’ file and then click on ‘Open’ button to get below prediction result



In above screen we can see uploaded file converted to binary grey image and then application predict malware contains in that file and this prediction result printed on image in white colour text and now test with another image



In above screen uploading 10.npy file and then click on ‘Open’ button to get below result



In above screen uploaded file contains ‘Worm Allaple.A’ malware and this prediction result printed on image in white colour text. Similarly you can upload other images and test