Meaningful Integration of Data, Analytics and Services

Grant Agreement No. 727721

Contract Duration: 40 months (1st November 2016 – 29th February 2020)

This project is funded by The European Union

H2020-SC1-2016-CNECT
SC1-PM-18-2016 - Big Data Supporting Public Health Policies

Deliverable 3.11

Synthetic Datasets 1

Circulation:	Public
Nature:	Other
Version #:	1.1
Issue Date:	02/07/2019
Responsible Partner(s):	Ulster University
Author(s):	Debbie Rankin, Michaela Black, Raymond Bond, Maurice Mulvenna, Jonathan Wallace, Brian Cleland
Status:	Final
Reviewed on:	17/07/2019
Reviewed by:	MIDAS Executive Board
Contractual Date of Delivery:	31/07/2019 (M33)

Grant Agreement No: 727721

Executive Board Document Sign Off

Role	Partner	Signature	Date
WP1 Lead	Ulster	Michaela Black	09/07/2019
WP2 Lead	SET	Paul Carlin	17/07/2019
WP3 Lead	VICOM	Gorka Epelde	09/07/2019
WP4 Lead	KU Leuven	Gorana Nikolic	10/07/2019
WP5 Lead	VTT	Juha Pajula	09/07/2019
WP6 Lead	DCU	Regina Connolly	09/07/2019
WP7 Lead	Ulster	Jonathan Wallace	17/07/2019
WP8 Lead	Ulster	Michaela Black	09/07/2019
Scientific-Technical Manager	Analytics Eng	Scott Fischaber	17/07/2019

Grant Agreement No: 727721

Abstract

Sharing data is often a risk in terms of security and privacy especially if the data is sensitive and related to a person’s health. Algorithms can be used to generate synthetic data from real data in order to share data that are considered more ‘privacy preserving’ and that increase the level of anonymity. In this task, we carry out experimental work to evaluate the validity of synthetic data as an alternative to real data when developing machine learning models. The evaluation metrics produced from machine learning models that are trained using synthetic data with metrics yielded from machine learning models that are trained using the corresponding real data are compared. Early findings indicate that synthetic data retains the properties and utility of the real data. A more extensive evaluation is required to prove this empirically, and to investigate disclosure risk.

Grant Agreement No: 727721

Copyright

- Ulster – University of Ulster (Project Coordinator) (UK)
- DCU – Dublin City University (Ireland)
- KU Leuven – Katholieke Universiteit Leuven (Belgium)
- VICOM – Fundación Centro De Tecnologías De Interacción Visual y Comunicaciones Vicomtech (Spain)
- UOULU – Oulun Yliopisto (University of Oulu) (Finland)
- ANALYTICS ENG – Analytics Engines Limited (UK)
- QUIN – Quintelligence D.O.O. (Slovenia)
- BSO – Regional Business Services Organisation (UK)
- DH – Department of Health (Public Health England) (UK)
- BIOEF – Fundación Vasca De Innovación E Investigación Sanitarias (Spain)
- VTT – Teknologian Tutkimuskeskus VTT Oy (Technical Research Centre of Finland Ltd.) (Finland)
- THL – Terveyden ja hyvinvoinnin laitos (National Institute for Health and Welfare) (Finland)
- SET – South Eastern Health & Social Care Trust (UK)
- IBM Ireland Ltd – IBM Ireland Limited (Ireland)
- ASU ABOR – Arizona State University (USA)

The MIDAS project is funded under the EC Horizon 2020 SC1- PMF-18 Big Data Supporting Public Health Policies

This document reflects only the author's views and the European Community is not liable for any use that might be made of the information contained herein. This document may not be copied, reproduced, or modified in whole or in part for any purpose without written permission from the MIDAS Consortium. In presence of such written permission, or when the circulation of the document is termed as “public”, an acknowledgement of the authors and of all applicable portions of the copyright notice must be clearly referenced. This document may change without prior notice.

Grant Agreement No: 727721

Document History

Version	Issue Date	Stage	Content and Changes
0.1	20/03/2019	Draft	Initial document draft outline
1.0	02/07/2019	Draft	V1 of document prepared for review
1.1	17/07/2019	Final	V1.1 final version with amendments made based on Executive Board suggestions.

Statement of Originality:

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Grant Agreement No: 727721

Executive Summary

Work Package:	WP 3
Work Package leader:	Fundación Centro De Tecnologías De Interacción Visual y Comunicaciones - Vicomtech (VICOM)
Task:	T 3.6 – Simulating Synthetic Data
Task leader:	University of Ulster

Synthetic data, also known as 'artificial data', is data that is simulated from real data using statistical models in order to represent the population in the original data whilst avoiding any divulgence of real, potentially personal, confidential and sensitive data. In the case of health-related data, this would ensure that actual patient records are not shared. Whilst they are somewhat representative, synthetic datasets avoid various governance and confidentiality issues since real patient or citizen records are not provided or disclosed. This task and first deliverable iteration (D3.11) involves the investigation and evaluation of synthetic data, initially utilising real, publicly available datasets. The synthetic versions of these datasets will be shared openly. The aim is to create synthetic data from the real population datasets that are made available in the MIDAS project in the next iteration of this task deliverable (D3.12). These will be shared openly if (upon rigorous evaluation) it can be proven that no disclosure risk remains in the synthetic data. As per the task in the Grant Agreement, this artificial data has been simulated using the SynthPop library inside the R programming environment. The synthetic datasets have been validated with the real data by analysing distributions, as well as by examining the performance of Machine Learning algorithms when applied to real data and comparing the results when the same algorithms are applied to the synthetic data. The evaluation metrics produced from machine learning models that are trained using synthetic data with metrics yielded from machine learning models that are trained using the corresponding real data have been compared. Early findings indicate that synthetic data retains the properties and utility of the real data. A more extensive evaluation is required to prove this empirically, and to investigate disclosure risk.

Grant Agreement No: 727721

Table of Contents

1 Introduction **8**

2 Methods **10**

2.1 Dataset Selection 10

2.2 Generating Synthetic Data 10

2.2.1 Synthetic Data with Numerical Data 11

2.2.2 Synthetic Data with Categorical Data 15

2.3 Machine Learning with Real and Synthetic Data 19

2.3.1 Machine Learning with the Breast Cancer Dataset 20

2.3.2 Machine Learning with the Nursery Dataset 21

3 Results **21**

3.1 Breast Cancer Dataset Results 21

3.2 Nursery Dataset Results 50

4 Conclusion **98**

5 Dissemination of Synthetic Datasets **100**

6 References **101**

7 Appendix A - Breast Cancer Dataset Decision Trees **103**

8 Appendix B - Breast Cancer Dataset Decision Trees Cross Comparison **110**

1 Introduction

The volume of data being generated every year is growing exponentially. A report from IBM in 2017 stated that 90% of the world's data was produced over the last two years and that over 2.5 quintillion bytes of data is generated every day (IBM Corporation, 2017). Data scientists are availing of this huge volume of data to solve real world problems for the greater good of society. Data science has already proven extremely valuable in areas such as fraud and risk detection, image analysis, speech recognition, internet search, and targeted marketing.

We know that data science also has the potential to vastly improve areas such as healthcare and cybersecurity and yet these improvements have not yet been fully realised. The reason may be in part related to an issue that faces many data scientists: the availability of data.

Privacy concerns over personal data, and in particular health care data, means that although the data exists, it is deemed too sensitive to be made available for public use, even in the case of serious research. Data sharing and data use demand careful governance, with the introduction of GDPR placing increasingly stringent guidelines on data management. Traditionally, data perturbation techniques have been applied to real data to modify and thus protect the data from disclosure prior to releasing it to users. Common methods include adding noise, data swapping, data masking, cell suppression, and stripping unique identifiers. However, such methods do not eliminate disclosure risk and can impact the utility of the data (Reiter, 2004a).

In the case of fraud detection, instances of fraud may be so rare that there is simply not enough data to allow data science techniques to be applied. Machine learning models require examples of fraud in order to learn, so that when they are faced with a previously unseen set of data they can accurately predict whether an observation should be classed as fraudulent or not fraudulent.

One way to overcome the issue of data availability is to use synthetic data as an alternative to real data. Synthetic data is generated from real data by using the underlying statistical properties of the real data to produce synthetic datasets that exhibit these same statistical properties.

Synthetic data was first proposed by Rubin (1993) and Little (1993). Raghunathan, Reiter and Rubin (2003) implemented and extended upon the approach, pioneering the parametric multiple imputation approach to synthetic data generation, a method based on the imputation of missing data but instead implemented for the purpose of synthesising data. A range of studies have since been published exemplifying this approach (Reiter, 2004, 2005a, 2005b, 2009 Reiter and Raghunathan, 2007, Reiter and Dreschler, 2010). Reiter (2005c) then introduced an alternative method of synthesising data through the non-parametric tree-based technique that utilises classification and regression trees (CART). Non-parametric methods have been shown to perform better in synthesising data compared to parametric methods. A more recent technique proposes generative modelling for synthetic data generation (Patki, Wedge and Veeramachaneni, 2016).

The aim of synthetic data is to enable data to be made publicly available, particularly for the purpose of serious research, that would typically be prevented from release, or be very slow to release, due to privacy and confidentiality concerns. The synthetic data should maintain the same statistical properties as the real data and should therefore be valid when used for inference.

Within the remit of the MIDAS project, data mining and machine learning techniques are being applied to real health-related data to derive knowledge that can be utilised within a healthcare policy decision making tool. This task seeks to ascertain whether synthetic data can preserve the hidden complex patterns that data mining can uncover from real data, and therefore whether it can be used as a valid alternative to real data when used in health care policy making. Some work has been completed in this area indicating promising results (Eno and Thompson, 2008, Heyburn et al., 2018). A good synthetic dataset should replace sensitive values and provide stronger guarantees of privacy and anonymity.

Synthetic data can be used in two ways:

1. To increase the size of a dataset, for times when a dataset is unbalanced due to the limited occurrence of an event.
2. To generate a full synthetic dataset that is representative of the original dataset, for times when data is not available due to its sensitive nature.

2 Methods

2.1 Dataset Selection

For initial experimentation, two publicly available health-related datasets have been selected. At this stage, datasets made available to the MIDAS project have not been utilised as we cannot fully guarantee that disclosure risk does not exist when the data are synthesised using the synthetic data generation techniques applied. Therefore it would be unsafe to make synthetic versions of such sensitive, confidential datasets openly available without further, more rigorous evaluation of disclosure risk. This will form part of the second version of this deliverable.

The first dataset analysed is the Breast Cancer Wisconsin dataset[1] (Mangasarian and Wolberg, 1990, Wolberg and Mangasarian, 1990, Mangasarian, Setiono and Wolberg, 1990, Bennett and Mangasarian, 1992). This dataset contains only numeric attributes, and contains 699 observations with ten attributes plus the class attribute. Each observation belongs to one of two classes: benign or malignant, represented as 2 and 4, respectively, in the dataset.

The second dataset analysed is the Nursery dataset[2] (Olave, Rajkovic and Bohanec, 1989, Zupan et al., 1997). This dataset contains only categorical attributes, and has 12,960 observations with eight attributes plus the class attribute. Each observation belongs to one of five classes: not_recom, recommend, very_recom, priority or spec_prior.

These datasets were selected to enable an analysis of synthetic data performance when applied to datasets of differing volume and attributes of differing data types, to determine whether these had an impact on analysis with machine learning algorithms.

2.2 Generating Synthetic Data

In this work we analyse and assess the performance of the parametric data synthesis technique of multiple imputation developed by Reiter (2004b), as well as the improved non-parametric tree-based synthesis technique that utilises CART (Reiter 2005c), as described in Section 1. The R package, Synthpop[3], developed by Nowak, Raab and Dibben (2016), provides a publicly available implementation of the

[1] https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original)
[2] https://archive.ics.uci.edu/ml/datasets/nursery
[3] https://cran.r-project.org/web/packages/synthpop/index.html

synthetic data generators. This implementation has been utilised in this experimental work.

2.2.1 Synthetic Data with Numerical Data

For each real dataset, five synthetic datasets were generated using the non-parametric method, and five synthetic datasets were generated using the parametric method. All parameters remained the same when generating these datasets. Multiple versions were synthesised using each approach to ensure experimental results were robust.

Attributes are synthesised sequentially in both the parametric and non-parametric methods. The first attribute to be synthesised in a dataset is a special case since it has no predictors from previously synthesised attributes in the dataset. The synthetic values for the first attribute are synthesised using a random sample from the original observed data, via the *Sample* method in Synthpop.

When synthesising attributes with the non-parametric method, Synthpop applies the *Cart* method, i.e. classification and regression trees. The *Cart* method can synthesise attributes of any data type. The *Cart* method is applied to all variables that have predictors, i.e. attributes prior to them in the sequence and draws from the conditional distributions fitted to the original data using CART models (Table 2.2.1).

When synthesising attributes with the parametric method, Synthpop applies synthesising methods based on the attribute data type. As the breast cancer dataset contains only numeric attributes, all are synthesised using normal linear regression via the *Norm Rank* function in Synthpop, except the first attribute that is synthesised using a random sample from the original data (Table 2.2.1).

Table 2.2.1 illustrates the model applied to each attribute in the Breast Cancer dataset when Non-Parametric and Parametric methods are applied, respectively.

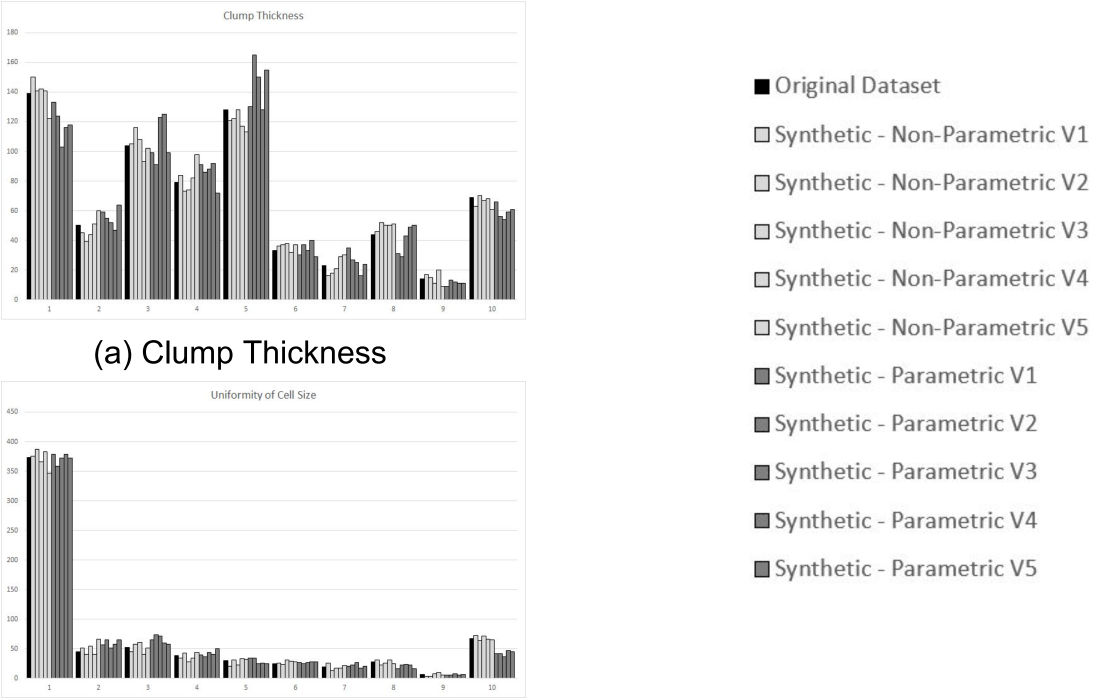
Table 2.2.1 Synthetic data generation models applied to each attribute in the Breast Cancer dataset when the Non-Parametric and Parametric synthesis methods are applied.

	Non-Parametric (CART)	**Parametric**
Sample Code Number	Sample	Sample
Clump Thickness	Cart	Norm Rank
Uniformity of Cell Size	Cart	Norm Rank
Uniformity of Cell Shape	Cart	Norm Rank
Marginal Adhesion	Cart	Norm Rank

Single Epithelial Cell Size	Cart	Norm Rank
Bare Nuclei	Cart	Norm Rank
Bland Chromatin	Cart	Norm Rank
Normal Nucleoli	Cart	Norm Rank
Mitoses	Cart	Norm Rank
Class	Cart	Norm Rank

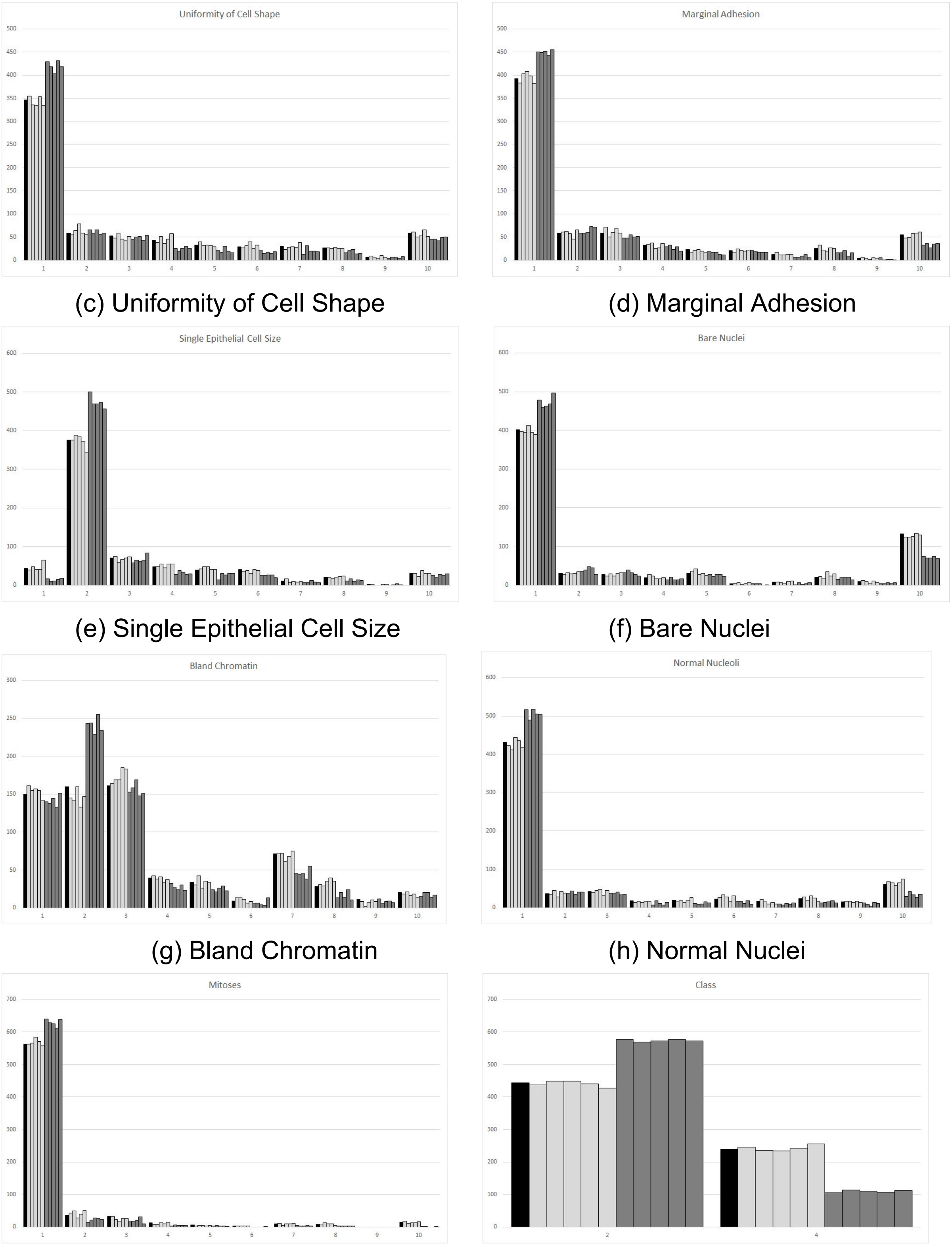
The Breast Cancer dataset contains 16 missing values in one attribute, *Bare Nuclei*. These missing values have been handled by removing the observations in which they occur. The dataset therefore has 683 observations remaining for synthesis. The missing values could have been imputed however, the impact of imputation is a separate investigation beyond the scope of this work.

Figure 2.2.1 illustrates the distributions of attributes from the original Breast Cancer dataset and the ten synthesised datasets, five generated with the non-parametric method and five with the parametric method. The SampleCode attribute is not included in these graphs as it is a unique identifier. We observe that the distributions of the synthetic Breast Cancer datasets generated using the non-parametric technique are very similar to the distribution of the original dataset. The distributions of attributes synthesised using the parametric method deviate slightly more from the original data with some attributes showing more deviation than others.



(a) Clump Thickness

(b) Uniformity of Cell Size



(c) Uniformity of Cell Shape

(d) Marginal Adhesion

(e) Single Epithelial Cell Size

(f) Bare Nuclei

(g) Bland Chromatin

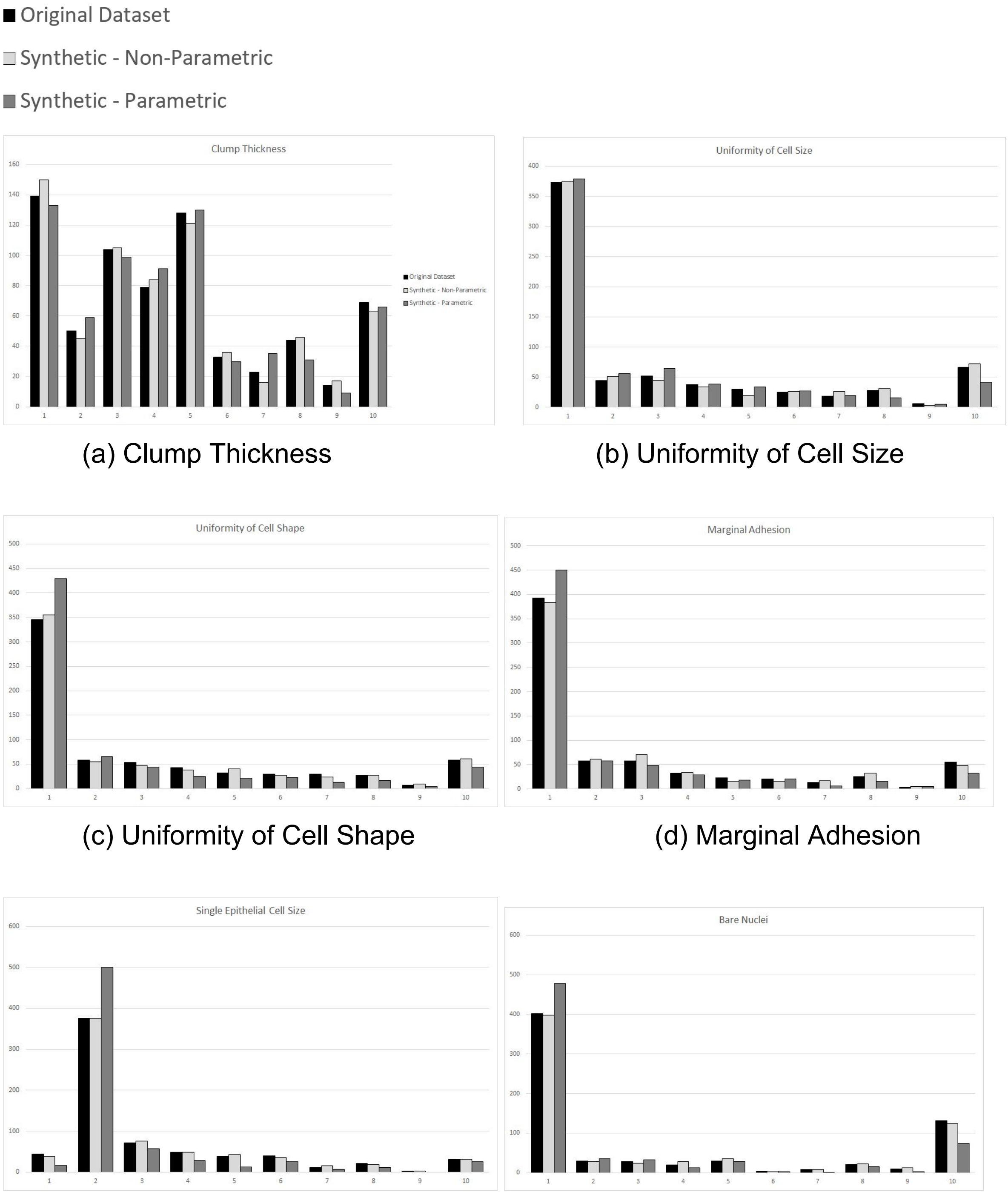
(h) Normal Nuclei

(i) Mitoses

(j) Class

Figure 2.2.1 (a)-(j) Distribution of each variable in the original dataset compared with those for the 10 synthetic datasets, 5 non-parametric, 5 parametric, for the Breast Cancer dataset

For better visibility the following set of graphs in Figure 2.2.2 shows a comparison of distributions for the original dataset and two synthetic datasets, one parametric (the first of the five datasets synthesised using the parametric method as shown in Figure 2.2.1 (a)-(j)) and one non-parametric (the first of the five synthesised using the non-parametric method as shown in Figure 2.2.1 (a)-(j)).



(a) Clump Thickness

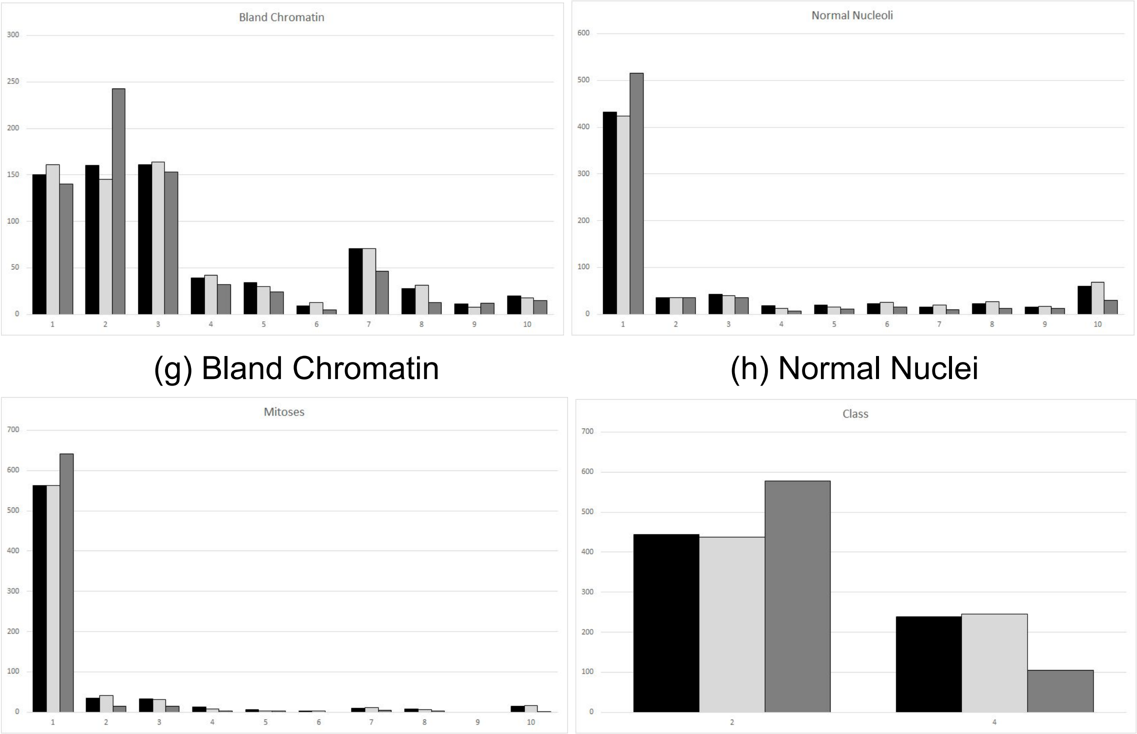
(b) Uniformity of Cell Size

(c) Uniformity of Cell Shape

(d) Marginal Adhesion

(e) Single Epithelial Cell Size

(f) Bare Nuclei



(g) Bland Chromatin

(h) Normal Nuclei

(i) Mitoses

(j) Class

Figure 2.2.2 (a)-(j) Distribution of each variable in the original dataset compared with those for 1 non-parametric synthetic dataset and 1 parametric synthetic dataset, for the Breast Cancer dataset

2.2.2 Synthetic Data with Categorical Data

In contrast, the Nursery dataset contains only categorical attributes. For synthesis with the non-parametric method the *Cart* method is applied to synthesise all attributes except the first, which is randomly sampled from the original data. For parametric synthesis, polytomous logistic regression is used to synthesise categorical variables with more than two levels via the *Polyreg* method in Synthpop, whilst logistic regression is applied to synthesise binary categorical variables via the *Logreg* method in Synthpop. Only one attribute, finance, has two possible values in the Nursery dataset. Table 2.2.3 illustrates the model applied to each attribute in the Nursery data when Non-Parametric and Parametric methods are applied.

Table 2.2.3 Synthetic data generation models applied to each attribute in the Nursery dataset when the Non-Parametric and Parametric synthesis methods are applied.

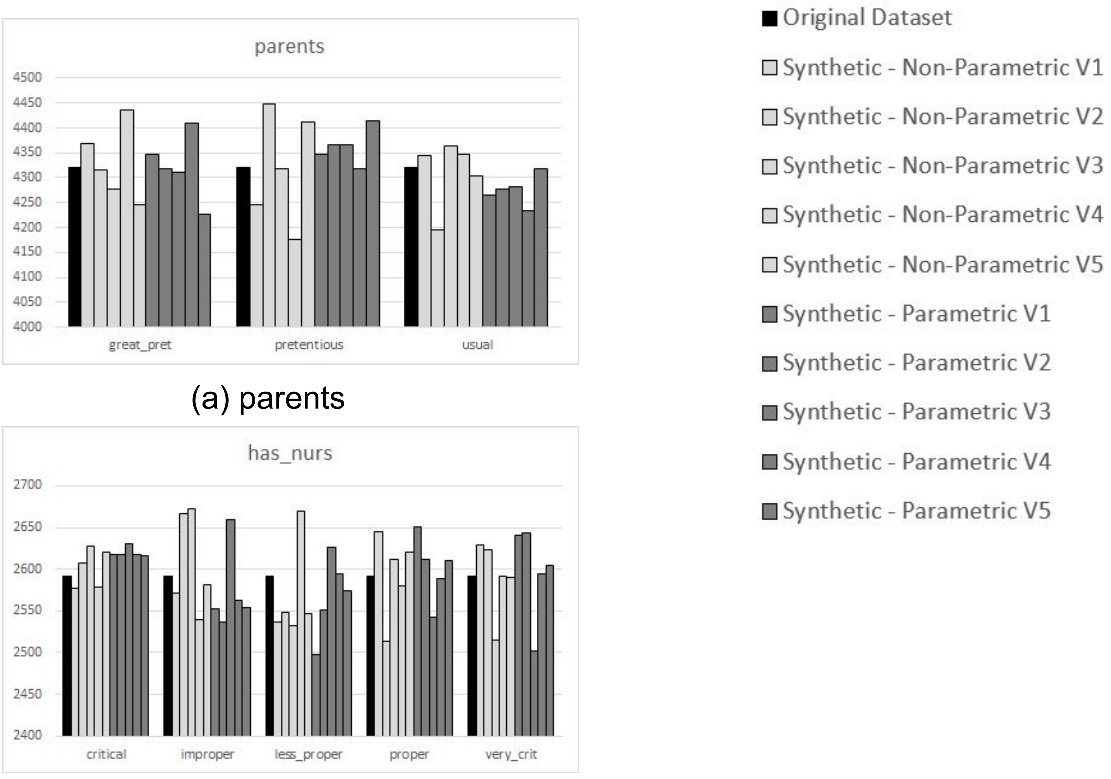
	Non-Parametric (CART)	**Parametric**
parents	Sample	Sample
has_nurs	Cart	Polyreg
form	Cart	Polyreg
children	Cart	Polyreg

housing	Cart	Polyreg
finance	Cart	Logreg
social	Cart	Polyreg
health	Cart	Polyreg
class	Cart	Polyreg

The Nursery dataset contains no missing values and so no records have been removed or imputed in this case.

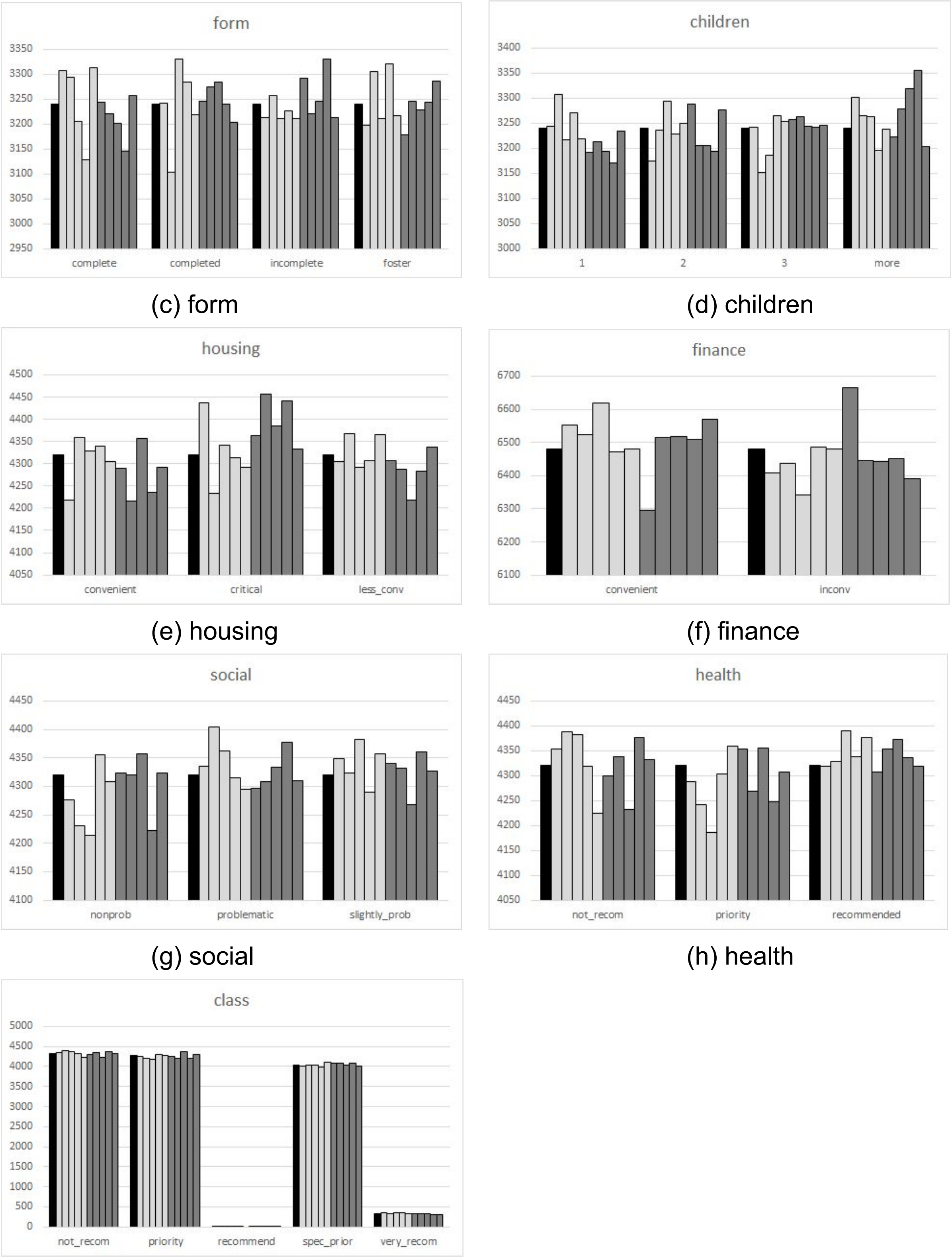
Figure 2.2.3 illustrates the distributions of attributes from the original Nursery dataset and the ten synthesised datasets, five generated with the non-parametric method and five with the parametric method.

We observe that the distributions of the synthetic Nursery datasets generated using the non-parametric and parametric methods, whilst similar to the distribution of the original dataset, do show a higher degree of deviation from the original compared with the synthesised numerical data from the Breast Cancer dataset. The difference in distributions of attributes synthesised using the parametric and non-parametric methods do not differ as much in the case of synthesised categorical attributes.



(a) parents

(b) has_nurs



(c) form

(d) children

(e) housing

(f) finance

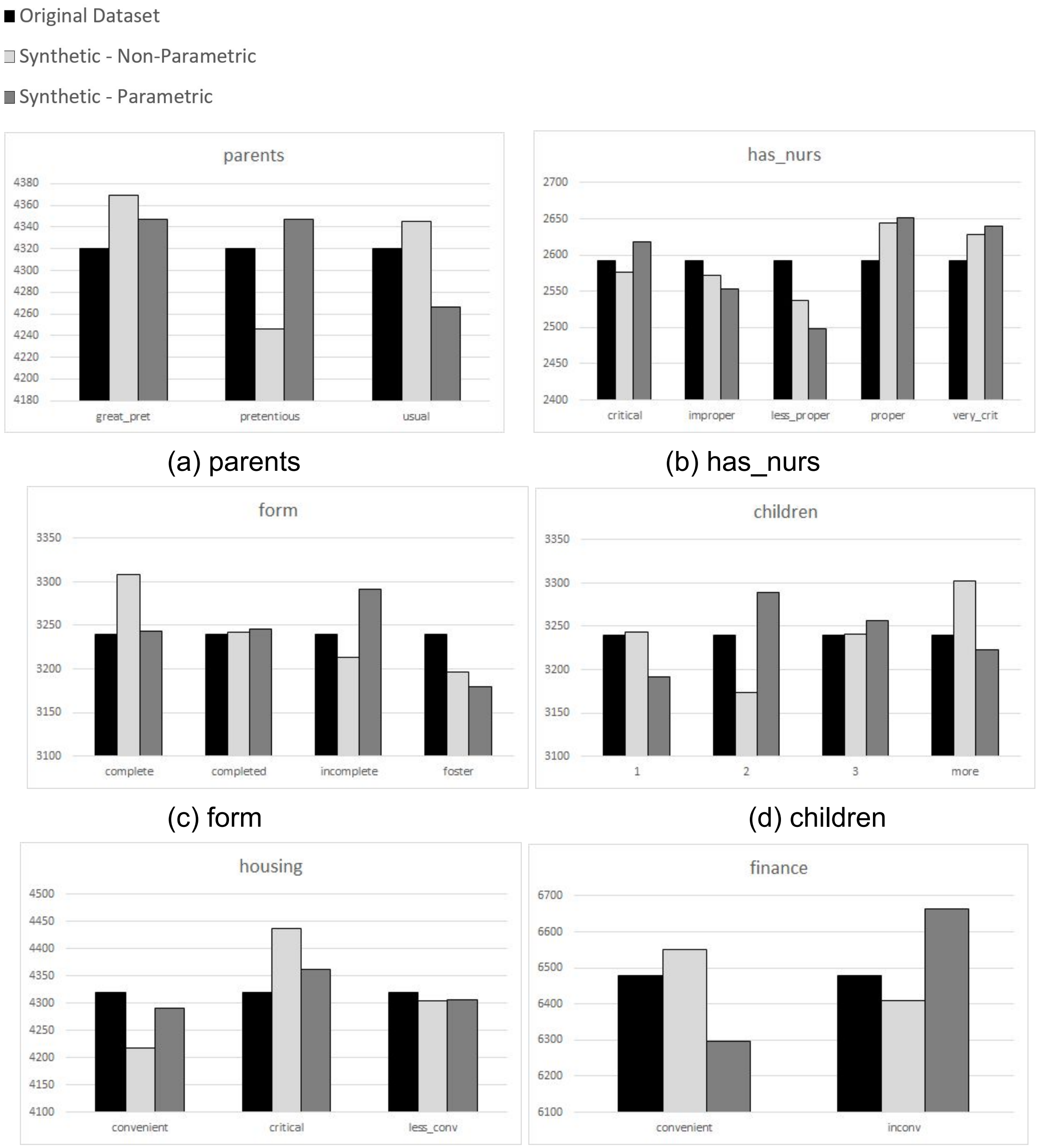
(g) social

(h) health

(i) class

Figure 2.2.3 (a)-(i) Distribution of each variable in the original dataset compared with those for the 10 synthetic datasets, 5 non-parametric, 5 parametric, for the Nursery dataset

For better visibility the following set of graphs in Figure 2.2.4 illustrates a comparison of distributions for the original dataset and two synthetic datasets, one parametric (the first of the five datasets synthesised using the parametric method as shown in Figure 2.2.3 (a)-(i)) and one non-parametric (the first of the five synthesised using the non-parametric method as shown in Figure 2.2.3 (a)-(i)).



(a) parents

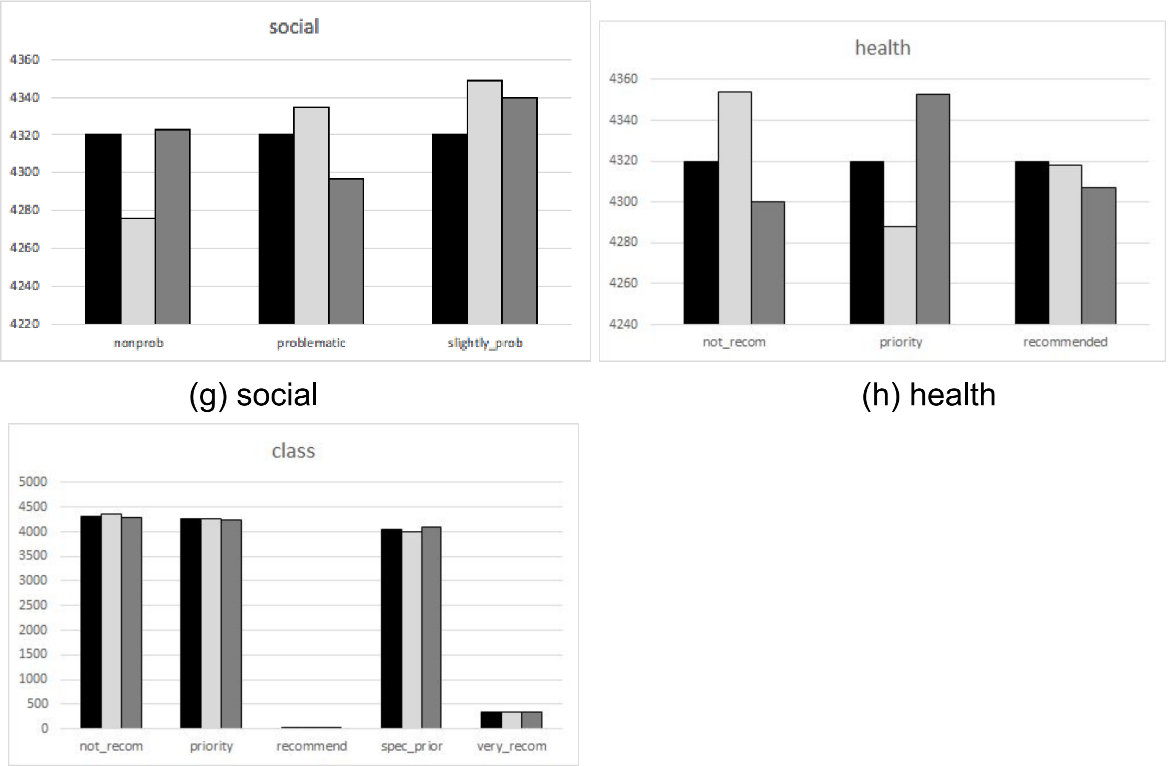
(b) has_nurs

(c) form

(d) children

(e) housing

(f) finance



(g) social

(h) health

(i) class

Figure 2.2.4 (a)-(i) Distribution of each variable in the original dataset compared with those for the 2 synthetic datasets, 1 non-parametric, 1 parametric, for the Nursery dataset

Overall, it is observed that the non-parametric synthesis method using CART performs better in synthesising numerical data compared with the parametric method. The difference in synthesis performance in categorical data between the parametric and non-parametric methods is negligible. The numerical Breast Cancer dataset is much smaller than the categorical Nursery dataset with 683 records compared with 12,960 records, respectively. Further work to determine if the size of the datasets has an impact on the performance of data synthesis is required. In addition, the significance of the difference between datasets must also be analysed.

2.3 Machine Learning with Real and Synthetic Data

To evaluate whether synthetic datasets can be used as a valid alternative to real datasets in machine learning, five different classification models were trained with the original Breast Cancer dataset, and the ten synthetic datasets described previously. The same methodology was also applied to the Nursery dataset.

2.3.1 Machine Learning with the Breast Cancer Dataset

The Breast Cancer dataset presents a binary classification problem. Therefore the range of models applied were: a Linear Classification model, a Decision Tree Classifier, a K-Nearest Neighbour Classifier, a Random Forest Classifier, and a Support Vector Machine Classifier.

This selection of algorithms were applied to determine how well each performed when trained with the original data compared with the synthetic data, with these classifiers ranging from simple to complex.

For training and testing, 10-fold cross validation (CV) was used, to reduce the risk of losing important patterns in the dataset and thus error induced from bias. The train/test split was 75/25.

The classifiers were implemented using Python's Scikit-Learn 0.21[4] machine learning library.

Linear classification was implemented using *SGDClassifier*, Stochastic Gradient Descent, a simple linear classifier, with loss="hinge", random_state=0 and all other parameters set to their defaults.

Decision tree classification was implemented using *DecisionTreeClassifier*, an optimised version of CART, with criterion="gini", max_depth=10 and random_state=0 and all other parameters set to their defaults.

The K-Nearest Neighbour classifier was implemented using *KNeighborsClassifier* with n_neighbors=10, weights='uniform', leaf_size=30, p=2, metric='minkowski', n_jobs=2 and all other parameters set to their defaults.

The Random Forest classifier was implemented using *RandomForestClassifier* with criterion="gini", max_depth=10, min_samples_split=2, n_estimators=10, random_state=1 and all other parameters set to their defaults.

The Support Vector Machine classifier was implemented using *SVC* with C=1.0, degree=3, kernel='rbf', probability=True, random_state=None and all other parameters set to their defaults.

[4] https://scikit-learn.org/stable/

2.3.2 Machine Learning with the Nursery Dataset

The Nursery dataset presents a multiclass classification problem. The Nursery dataset contains only categorical data. Classifiers in Scikit-Learn cannot readily handle categorical data. Therefore the categorical attributes were transformed into indicator attributes using one-hot encoding, where each categorical feature becomes an array whose size is the number of possible choices for that feature.

The same range of models were applied to the Nursery dataset as were applied to the Breast Cancer dataset with the same parameters (as described in Section 2.3.1): a Linear Classification model, a Decision Tree Classifier, a K-Nearest Neighbour Classifier, a Random Forest Classifier, and a Support Vector Machine Classifier.The classifiers were implemented using Python's Scikit-Learn machine learning library.

This selection of algorithms were applied to determine how well each performed when trained with the original data compared with the synthetic data, with these classifiers ranging from simple to complex.

Again, for training and testing, 10-fold cross validation (CV) was used. The train/test split was 75/25.

3 Results

3.1 Breast Cancer Dataset Results

To compare the performance of each model after being trained with the original and synthetic datasets, a variety of evaluation metrics were used. Firstly, the accuracy of each model was computed. Table 3.1.1 and figure 3.1.1 illustrate the accuracy of each of the five classification models after being trained by the original dataset and the ten synthetic datasets (five non-parametric and five parametric) and tested using 10 cross-fold validation with a 75/25 train/test split.

Table 3.1.1 Accuracy scores achieved by each model trained by each Breast Cancer dataset

Dataset	Linear Model	Decision Tree	KNN	Random Forest	SVM
Original	0.964	0.946	0.969	0.968	**0.973**
Synthetic Non-Parametric V1	0.946	0.953	0.947	**0.957**	0.949
Synthetic Non-Parametric V2	0.936	0.927	0.953	0.943	**0.961**

Synthetic Non-Parametric V3	0.956	0.953	0.960	0.958	**0.964**
Synthetic Non-Parametric V4	0.957	0.949	0.956	0.957	**0.959**
Synthetic Non-Parametric V5	0.935	0.950	0.958	0.956	**0.959**
Synthetic Parametric V1	0.939	0.930	0.956	0.952	**0.958**
Synthetic Parametric V2	0.951	0.943	0.954	0.957	**0.962**
Synthetic Parametric V3	0.963	0.949	0.962	0.965	**0.966**
Synthetic Parametric V4	0.971	0.958	0.970	0.964	**0.975**
Synthetic Parametric V5	0.950	0.930	0.958	0.953	**0.961**

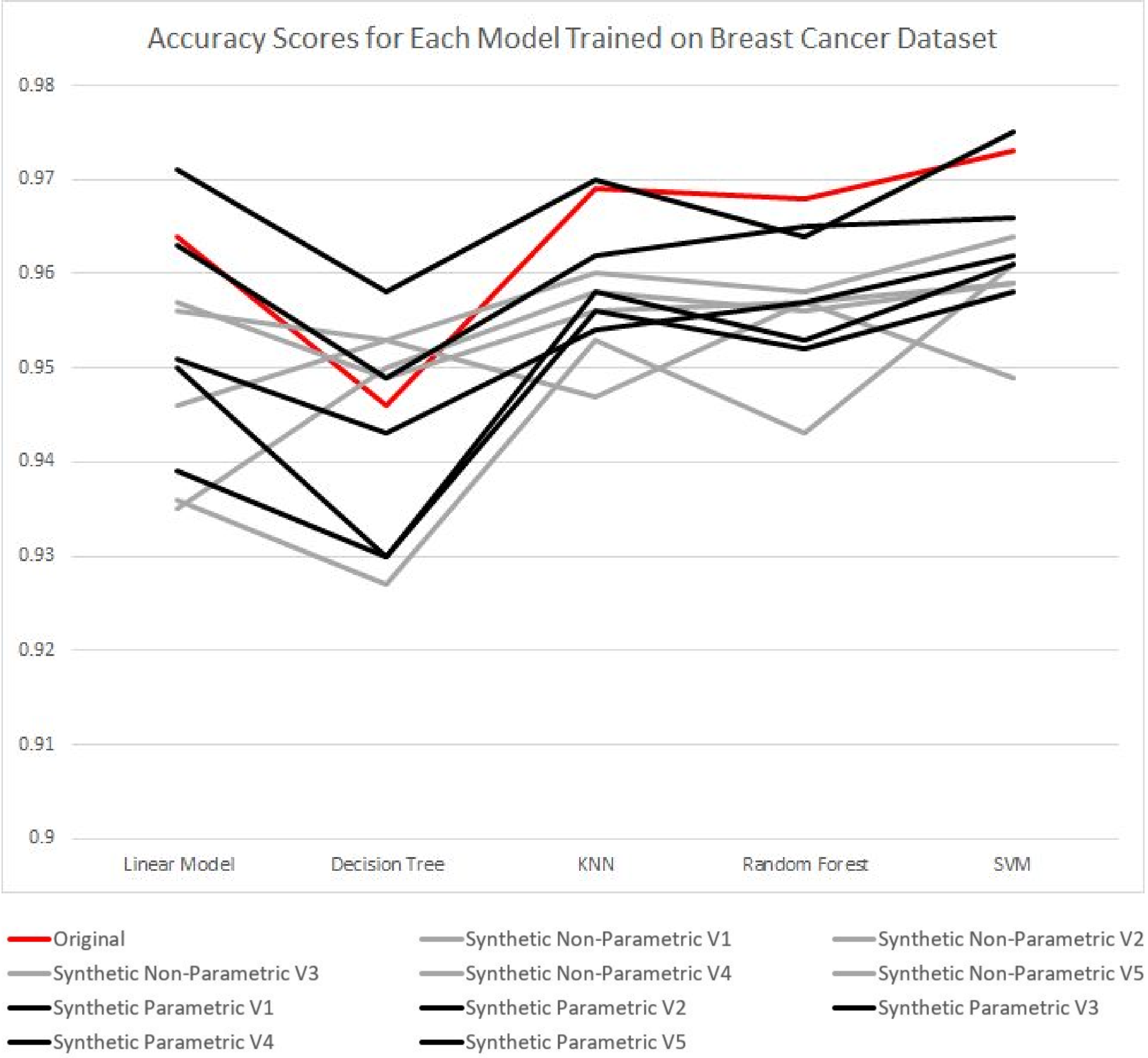


Figure 3.1.1 Accuracy scores achieved by each model as trained by each Breast Cancer dataset

We observe that all models perform well on the original and synthetic datasets. The minimum accuracy calculated was 0.927 from a Decision Tree applied to one of the five non-parametric synthetic datasets. Whilst lower than the others, this is still a very good rate of accuracy. The maximum accuracy calculated was 0.975 from an SVM applied to one of the five parametric datasets. The most accurate model overall is SVM followed by KNN, however all models perform well and the performance difference between the real dataset and the synthetic datasets generated using both parametric and non-parametric methods is negligible.

In addition to accuracy, precision scores, recall scores and the F1 measure were computed to gain a full understanding of how the models performed on real versus synthetic data. Tables 3.1.2-3.1.4 and figures 3.1.2-3.1.4 illustrate the precision, recall and F1 measures, respectively, for each of the five classification models after being trained by the original dataset and the ten synthetic datasets. We observe that precision and F1 scores for each model and for each dataset offer similar insights into model performance as the accuracy score. Recall scores have a higher degree of similarity across each model when applied to the same dataset, however models trained with synthetic data generated using parametric methods obtain better recall scores when compared with the recall scores of models trained with data generated using non-parametric methods. Precision and F1 scores are highest for the original, real dataset, whilst in contrast, recall scores are lower for the real dataset compared with the synthetic datasets. Visualisations of the decision trees trained from this data are provided in Appendix A for reference.

Table 3.1.2 Comparison of precision scores achieved by each model as trained by each dataset

Dataset	**Linear Model**	**Decision Tree**	**KNN**	**Random Forest**	**SVM**
Original	0.963	0.946	0.969	0.966	**0.971**
Synthetic Non-Parametric V1	0.946	0.952	0.951	**0.955**	0.950
Synthetic Non-Parametric V2	0.939	0.923	0.950	0.941	**0.957**
Synthetic Non-Parametric V3	0.956	0.950	0.962	0.958	**0.965**
Synthetic Non-Parametric V4	0.951	0.943	0.952	0.952	**0.954**
Synthetic Non-Parametric V5	0.935	0.948	0.953	0.952	**0.955**
Synthetic Parametric V1	0.917	0.870	0.957	0.933	**0.961**
Synthetic Parametric V2	0.928	0.909	0.951	0.945	**0.956**
Synthetic Parametric V3	0.932	0.909	0.951	0.945	**0.953**
Synthetic Parametric V4	0.948	0.929	0.967	0.956	**0.970**
Synthetic Parametric V5	0.926	0.870	**0.948**	0.938	**0.948**

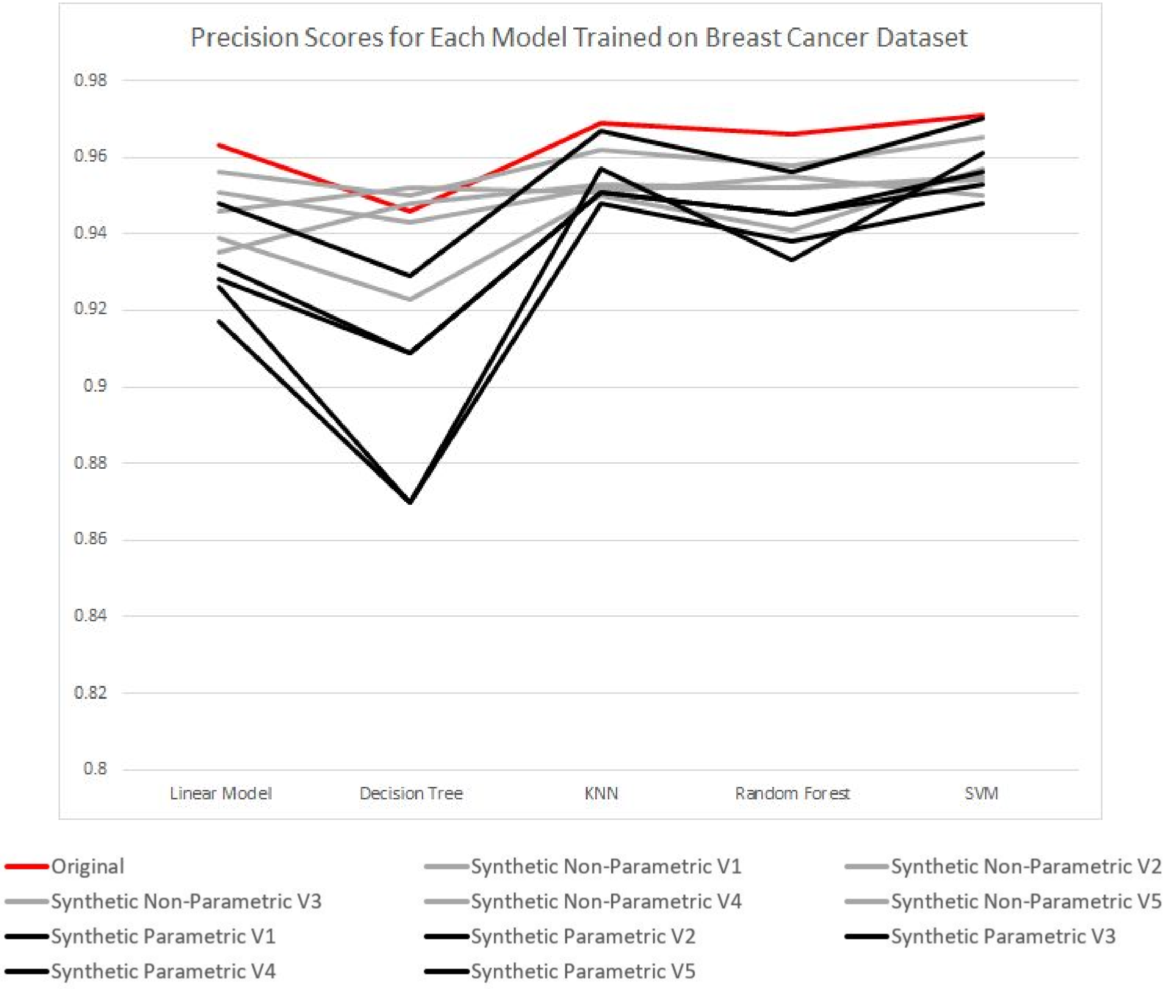


Figure 3.1.2 Comparison of precision scores achieved by each model as trained by each dataset

Table 3.1.3 Comparison of recall scores achieved by each model as trained by each dataset

Dataset	Linear Model	Decision Tree	KNN	Random Forest	SVM
Original	0.870	0.874	0.876	**0.886**	0.883
Synthetic Non-Parametric V1	**0.915**	0.897	0.888	0.906	0.912
Synthetic Non-Parametric V2	**0.942**	0.910	0.907	0.927	0.922
Synthetic Non-Parametric V3	**0.947**	0.918	0.921	0.910	0.937
Synthetic Non-Parametric V4	0.895	0.878	0.893	0.886	**0.907**
Synthetic Non-Parametric V5	0.960	0.938	0.965	0.964	**0.969**
Synthetic Parametric V1	0.941	0.950	0.943	**0.954**	0.945
Synthetic Parametric V2	0.926	0.919	0.948	0.937	**0.957**
Synthetic Parametric V3	0.950	0.947	0.955	0.953	**0.959**
Synthetic Parametric V4	0.953	0.943	0.950	0.953	**0.954**
Synthetic Parametric V5	0.929	0.946	0.954	0.952	**0.956**

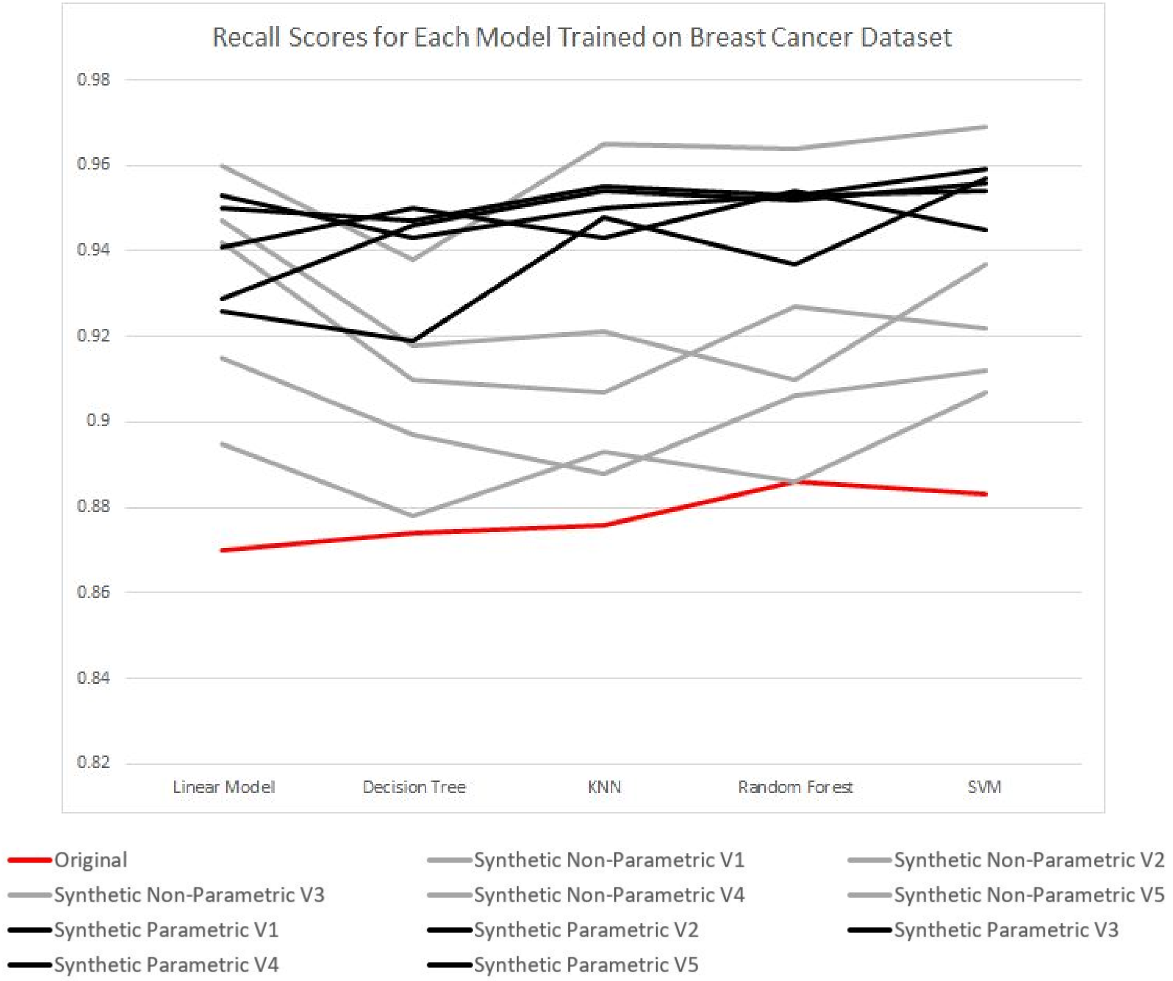


Figure 3.1.3 Comparison of recall scores achieved by each model as trained by each dataset

Table 3.1.4 Comparison of f1 scores achieved by each model as trained by each dataset

Dataset	Linear Model	Decision Tree	KNN	Random Forest	SVM
Original	0.960	0.938	0.965	0.964	**0.969**
Synthetic Non-Parametric V1	0.941	0.950	0.943	**0.954**	0.945
Synthetic Non-Parametric V2	0.926	0.919	0.948	0.937	**0.957**
Synthetic Non-Parametric V3	0.950	0.947	0.955	0.953	**0.959**
Synthetic Non-Parametric V4	0.953	0.943	0.950	0.953	**0.954**
Synthetic Non-Parametric V5	0.929	0.946	0.954	0.952	**0.956**
Synthetic Parametric V1	0.880	0.871	0.909	0.906	**0.915**
Synthetic Parametric V2	0.916	0.902	0.915	0.923	**0.931**
Synthetic Parametric V3	0.935	0.909	0.926	0.935	**0.936**
Synthetic Parametric V4	0.946	0.922	0.942	0.930	**0.952**
Synthetic Parametric V5	0.904	0.872	0.916	0.908	**0.924**

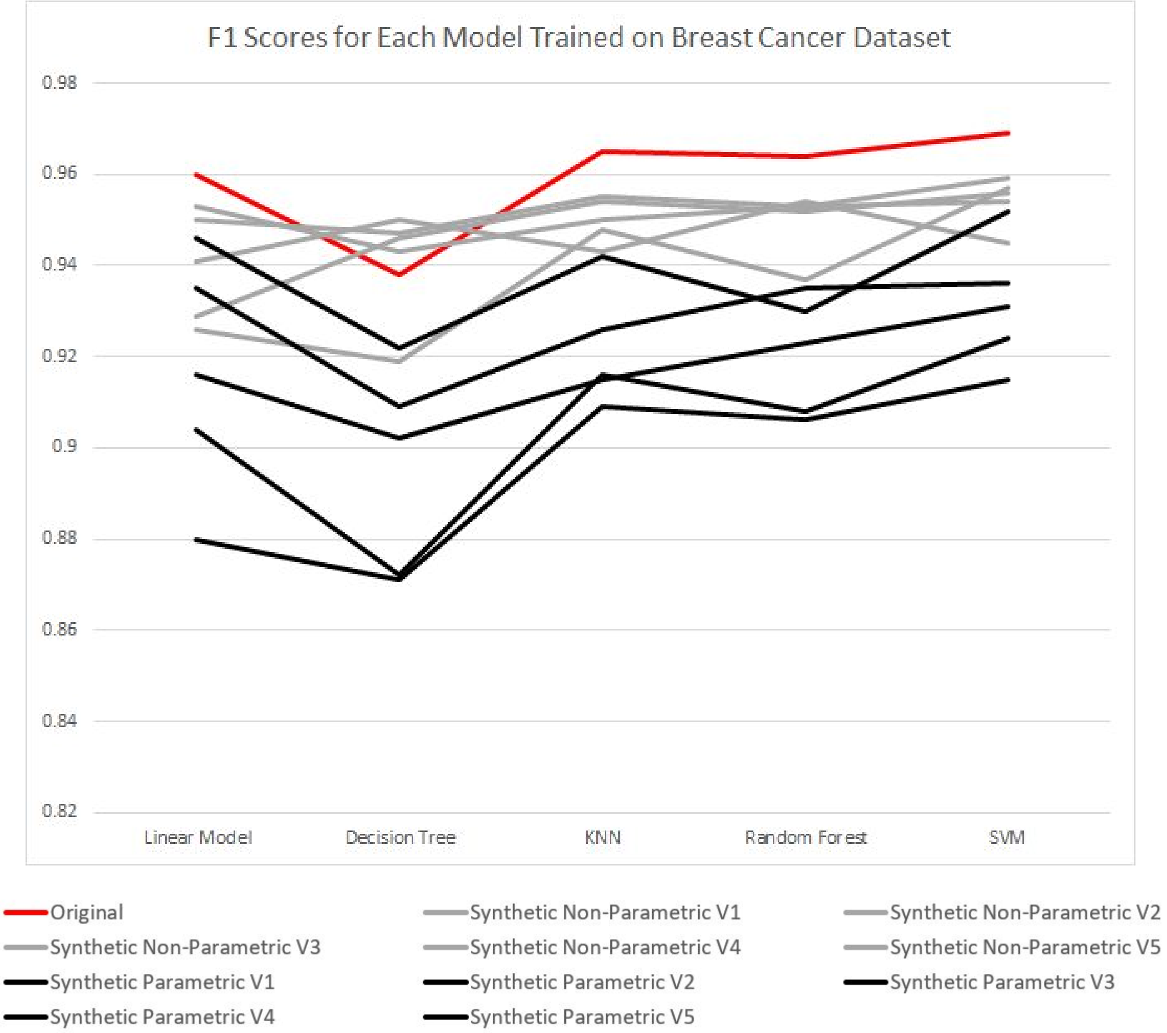
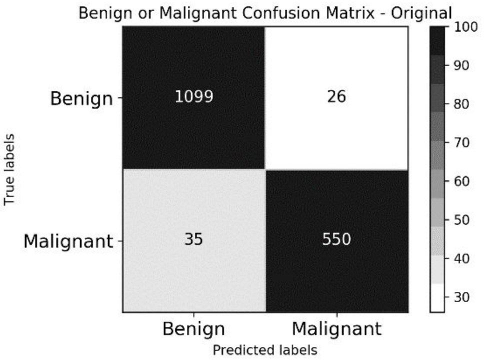
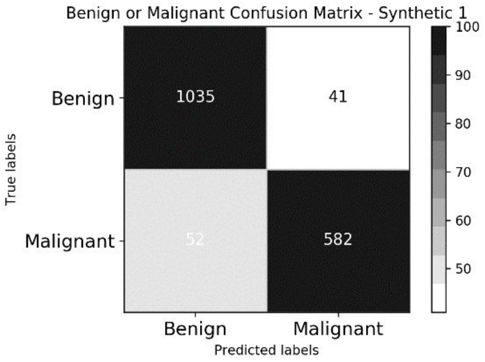
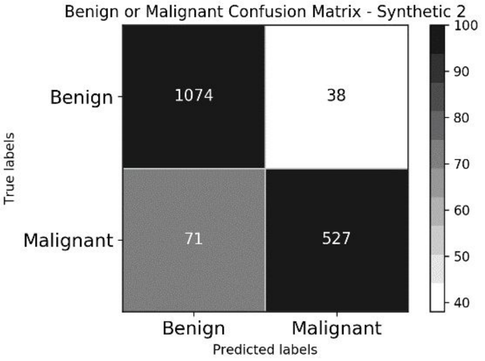


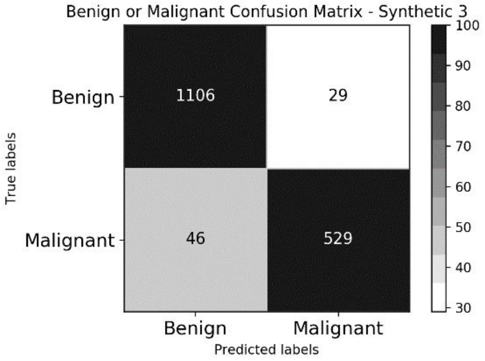
Figure 3.1.4 Comparison of f1 scores achieved by each model as trained by each dataset

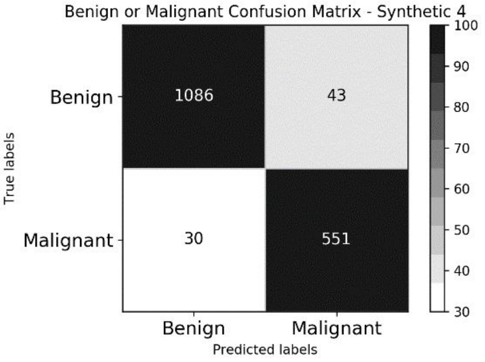
Although precision, accuracy, recall and F1 measures are summaries of the confusion matrix in some form, it is still beneficial to separate out the decisions made by the model to show where one class is being misclassified for another (false positives and false negatives). The confusion matrices for the performance of each of the five classifiers, trained on each of the eleven datasets (original, 5 synthetic non-parametric and 5 synthetic parametric) are shown in Figure 3.1.5-3.1.9 for the Linear model, Decision Tree model, KNN model, Random Forest model and SVM model, respectively. In all models and for each of the datasets, original and synthetic, the true positives and true negatives are high, however false positives and false negatives still occur. If we wanted to utilise data such as that in this breast cancer dataset to produce a classification model that can determine at the patient level, whether a tumour is benign or malignant, then the presence of false positives and false negatives is a concern. False negatives are of particular concern as, in this case, a tumour may be falsely classified as benign when it is in fact malignant. It should be noted however, that this issue exists in the models created using both the real data and synthetic data, and therefore the issue cannot be confirmed to be related to the synthetic data as it exists in the real data too. With some fine tuning of the models, false positives and false negatives could potentially be reduced in models created from both real and synthetic data.

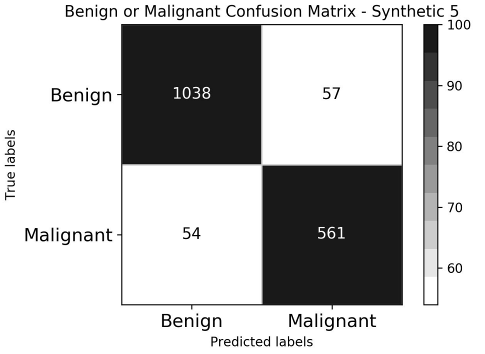












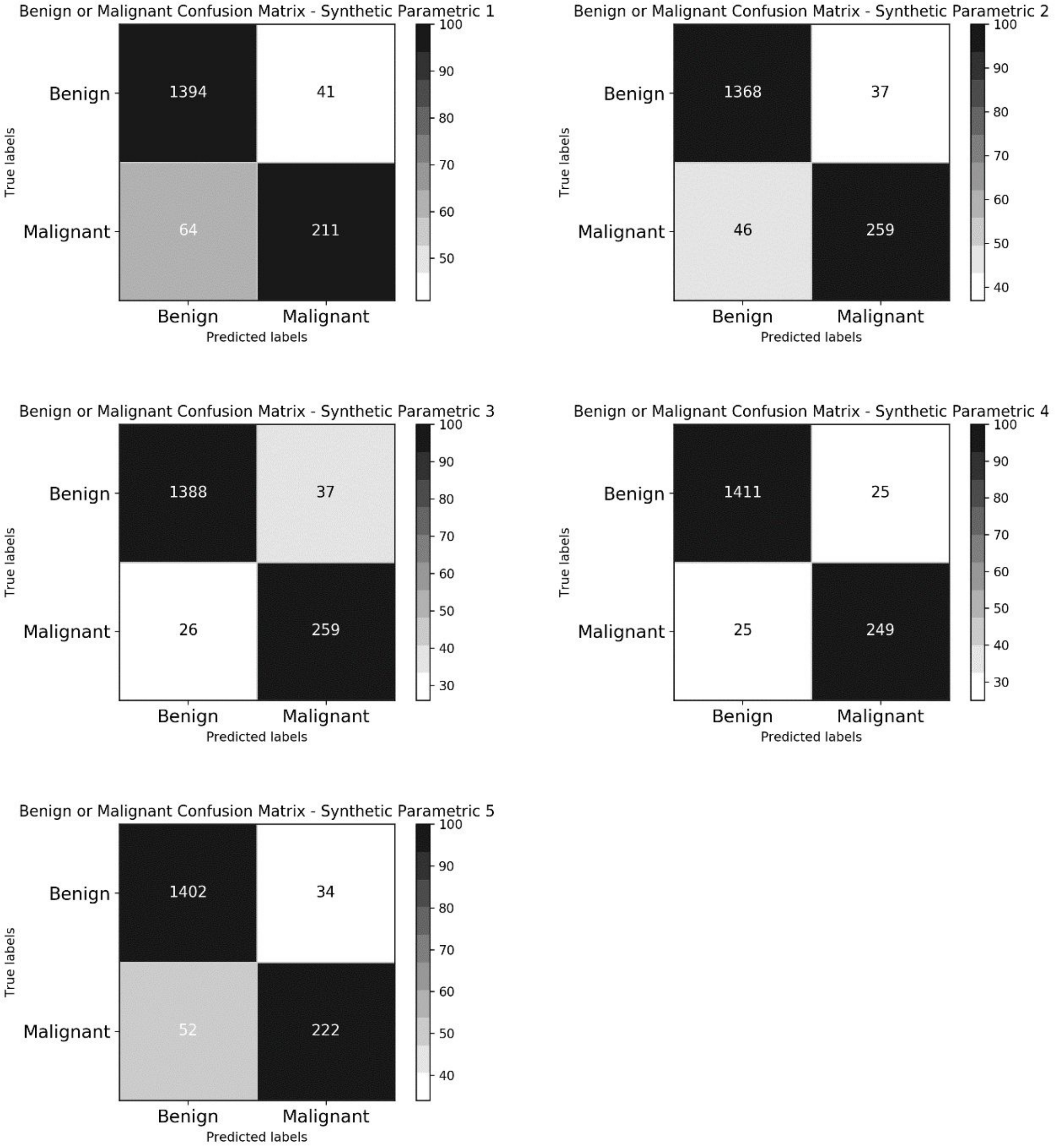
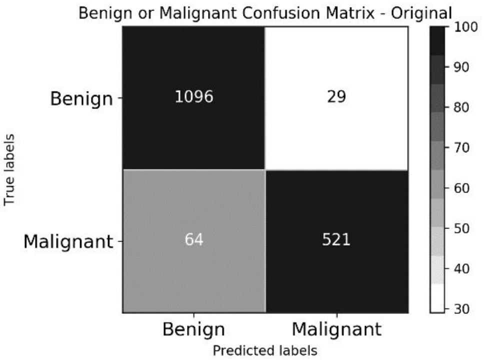
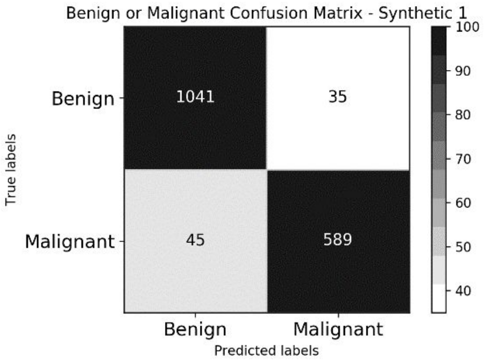
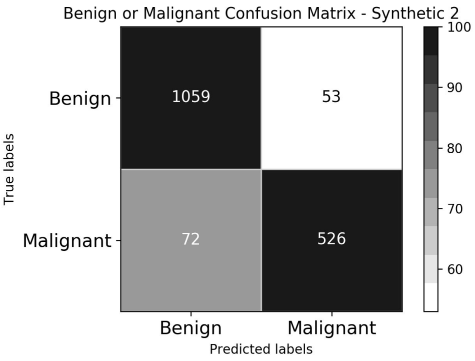
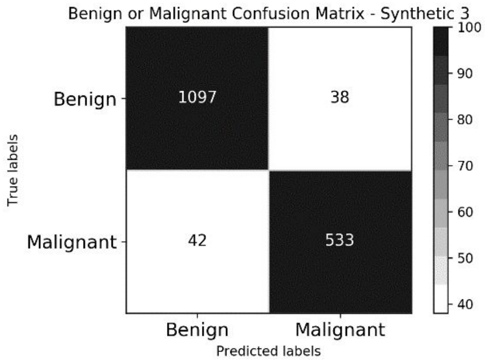


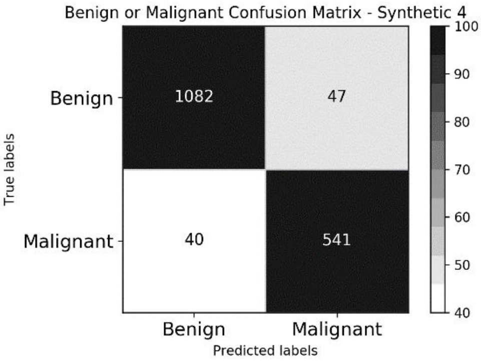
Figure 3.1.5 Confusion Matrices for the Linear Model when applied to each of the 11 datasets (1 original and 10 synthetic)

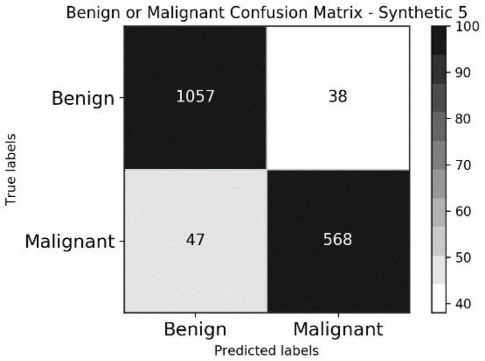

Benign or Malignant Confusion Matrix - Original

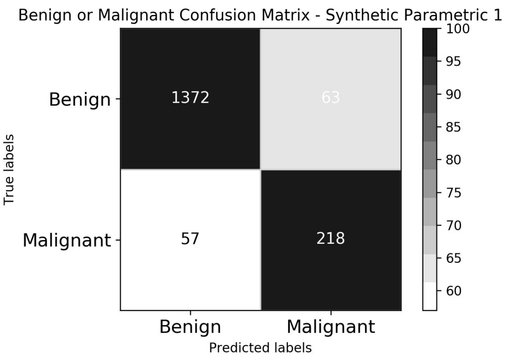

Benign or Malignant Confusion Matrix - Synthetic 1

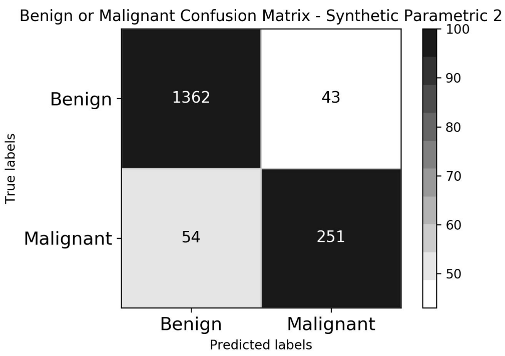

Benign or Malignant Confusion Matrix - Synthetic 2


Benign or Malignant Confusion Matrix - Synthetic 3


Benign or Malignant Confusion Matrix - Synthetic 4


Benign or Malignant Confusion Matrix - Synthetic 5


Benign or Malignant Confusion Matrix - Synthetic Parametric 1


Benign or Malignant Confusion Matrix - Synthetic Parametric 2

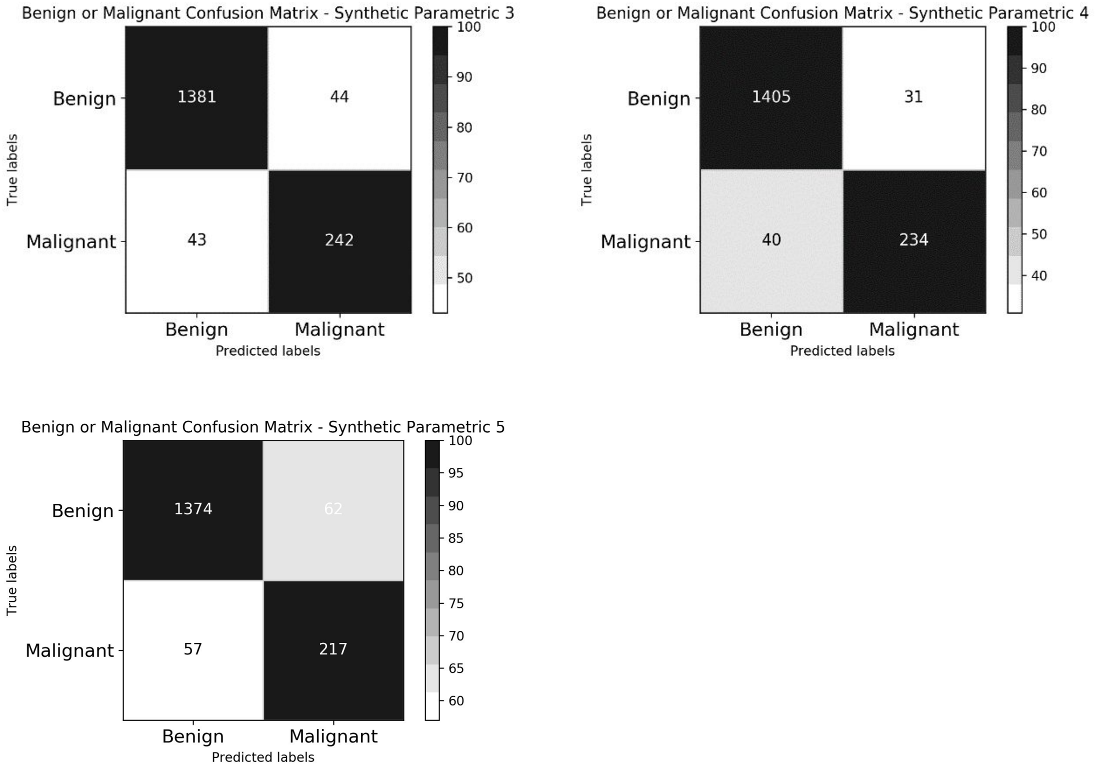
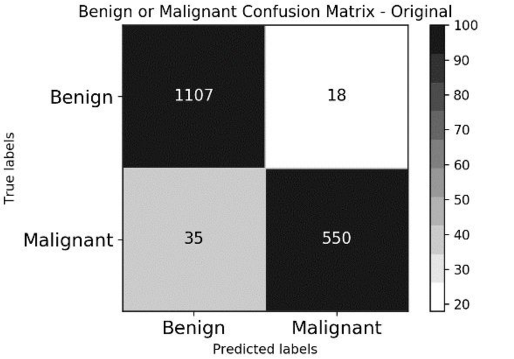
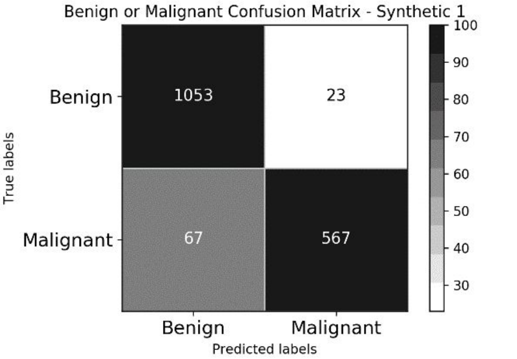
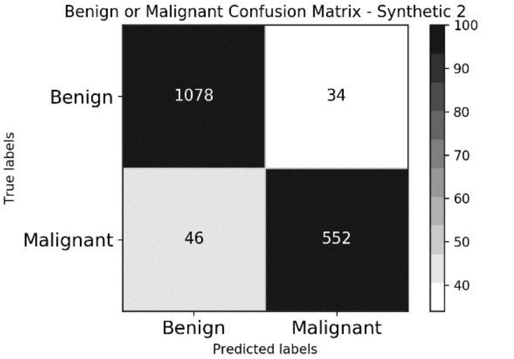
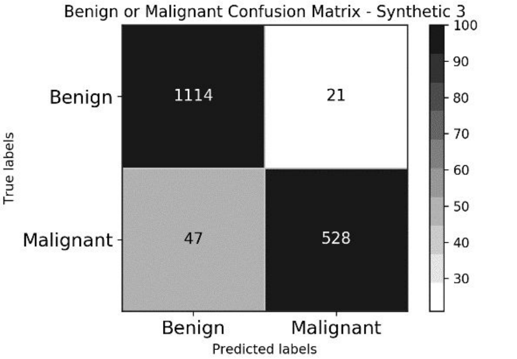


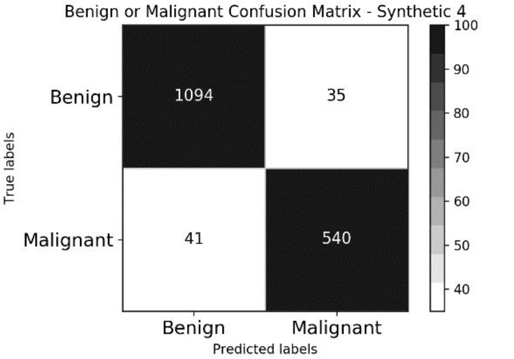
Figure 3.1.6 Confusion Matrices for the Decision Tree Model when applied to each of the 11 datasets (1 original and 10 synthetic)

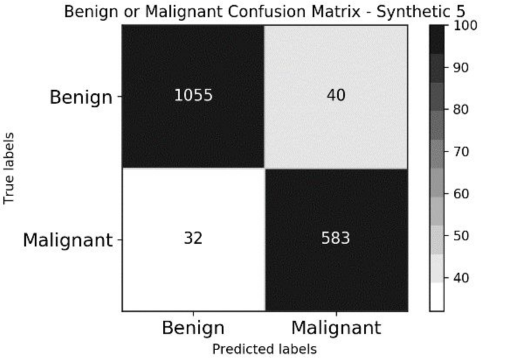

Benign or Malignant Confusion Matrix - Original
True labels
Benign
Malignant
1107
18
35
550
Benign
Malignant
Predicted labels
100
90
80
70
60
50
40
30
20

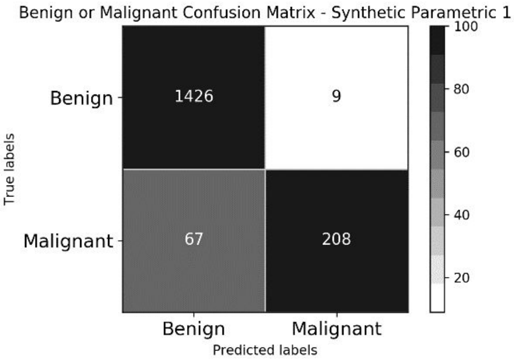

Benign or Malignant Confusion Matrix - Synthetic 1
True labels
Benign
Malignant
1053
23
67
567
Benign
Malignant
Predicted labels
100
90
80
70
60
50
40
30

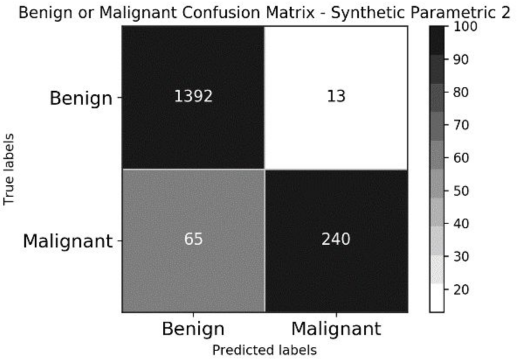

Benign or Malignant Confusion Matrix - Synthetic 2
True labels
Benign
Malignant
1078
34
46
552
Benign
Malignant
Predicted labels
100
90
80
70
60
50
40


Benign or Malignant Confusion Matrix - Synthetic 3
True labels
Benign
Malignant
1114
21
47
528
Benign
Malignant
Predicted labels
100
90
80
70
60
50
40
30


Benign or Malignant Confusion Matrix - Synthetic 4
True labels
Benign
Malignant
1094
35
41
540
Benign
Malignant
Predicted labels
100
90
80
70
60
50
40


Benign or Malignant Confusion Matrix - Synthetic 5
True labels
Benign
Malignant
1055
40
32
583
Benign
Malignant
Predicted labels
100
90
80
70
60
50
40


Benign or Malignant Confusion Matrix - Synthetic Parametric 1
True labels
Benign
Malignant
1426
9
67
208
Benign
Malignant
Predicted labels
100
80
60
40
20


Benign or Malignant Confusion Matrix - Synthetic Parametric 2
True labels
Benign
Malignant
1392
13
65
240
Benign
Malignant
Predicted labels
100
90
80
70
60
50
40
30
20

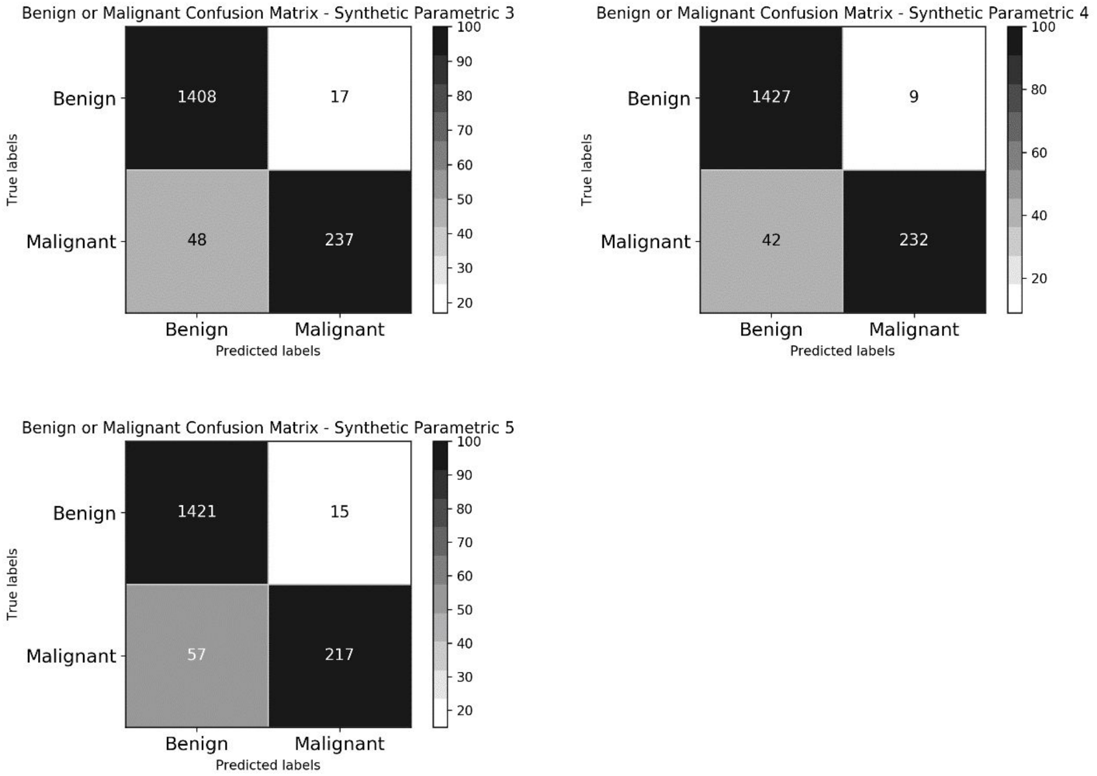
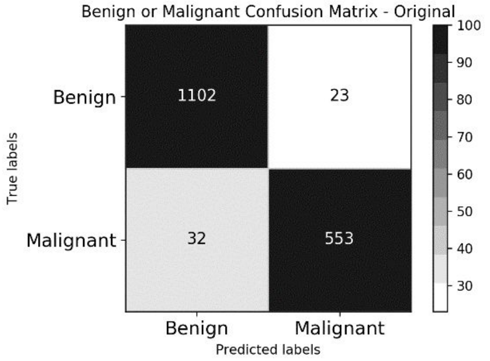
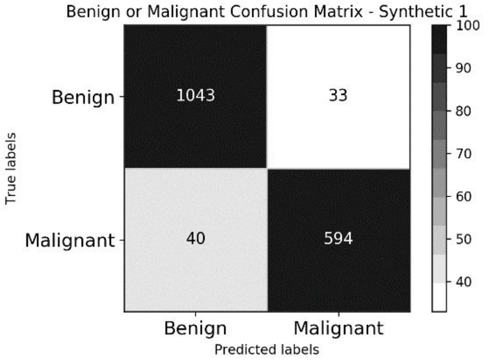
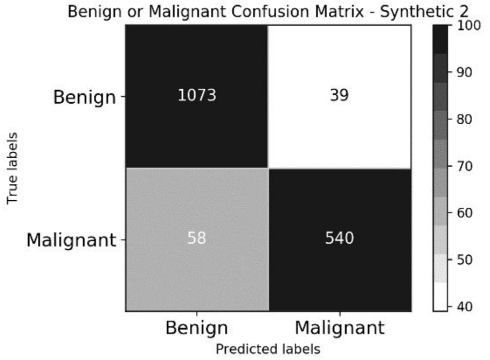
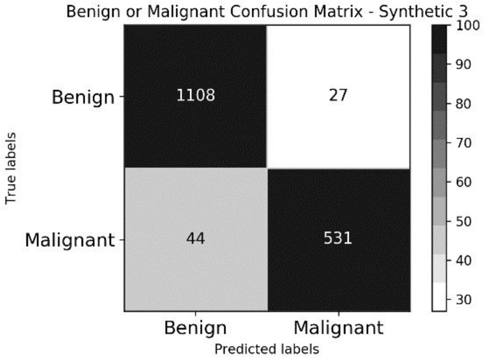


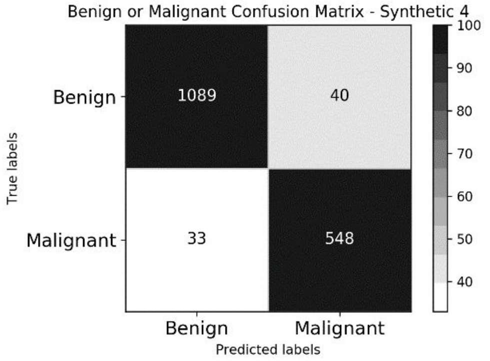
Figure 3.1.7 Confusion Matrices for the KNN Model when applied to each of the 11 datasets (1 original and 10 synthetic)

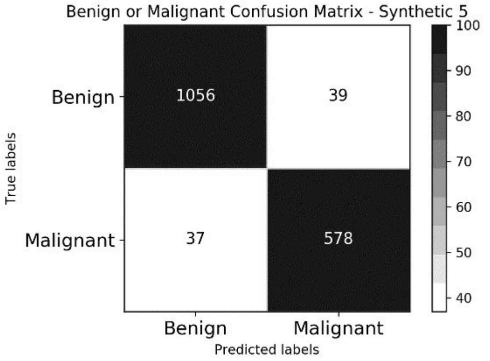

Benign or Malignant Confusion Matrix - Original
Benign
Malignant
True labels
1102
23
32
553
Benign
Malignant
Predicted labels
100
90
80
70
60
50
40
30

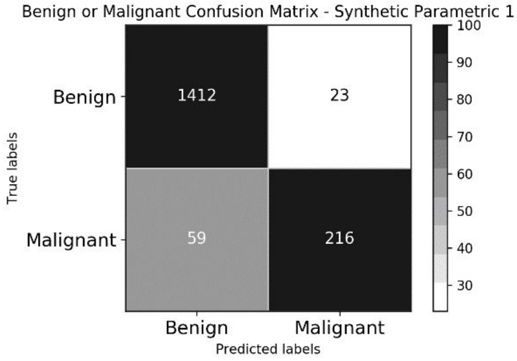

Benign or Malignant Confusion Matrix - Synthetic 1
Benign
Malignant
True labels
1043
33
40
594
Benign
Malignant
Predicted labels
100
90
80
70
60
50
40

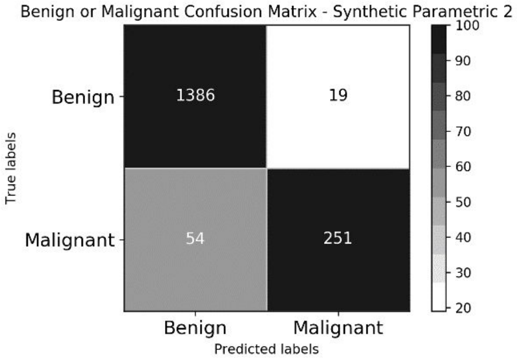

Benign or Malignant Confusion Matrix - Synthetic 2
Benign
Malignant
True labels
1073
39
58
540
Benign
Malignant
Predicted labels
100
90
80
70
60
50
40


Benign or Malignant Confusion Matrix - Synthetic 3
Benign
Malignant
True labels
1108
27
44
531
Benign
Malignant
Predicted labels
100
90
80
70
60
50
40
30


Benign or Malignant Confusion Matrix - Synthetic 4
Benign
Malignant
True labels
1089
40
33
548
Benign
Malignant
Predicted labels
100
90
80
70
60
50
40


Benign or Malignant Confusion Matrix - Synthetic 5
Benign
Malignant
True labels
1056
39
37
578
Benign
Malignant
Predicted labels
100
90
80
70
60
50
40


Benign or Malignant Confusion Matrix - Synthetic Parametric 1
Benign
Malignant
True labels
1412
23
59
216
Benign
Malignant
Predicted labels
100
90
80
70
60
50
40
30


Benign or Malignant Confusion Matrix - Synthetic Parametric 2
Benign
Malignant
True labels
1386
19
54
251
Benign
Malignant
Predicted labels
100
90
80
70
60
50
40
30
20

Grant Agreement No: 727721

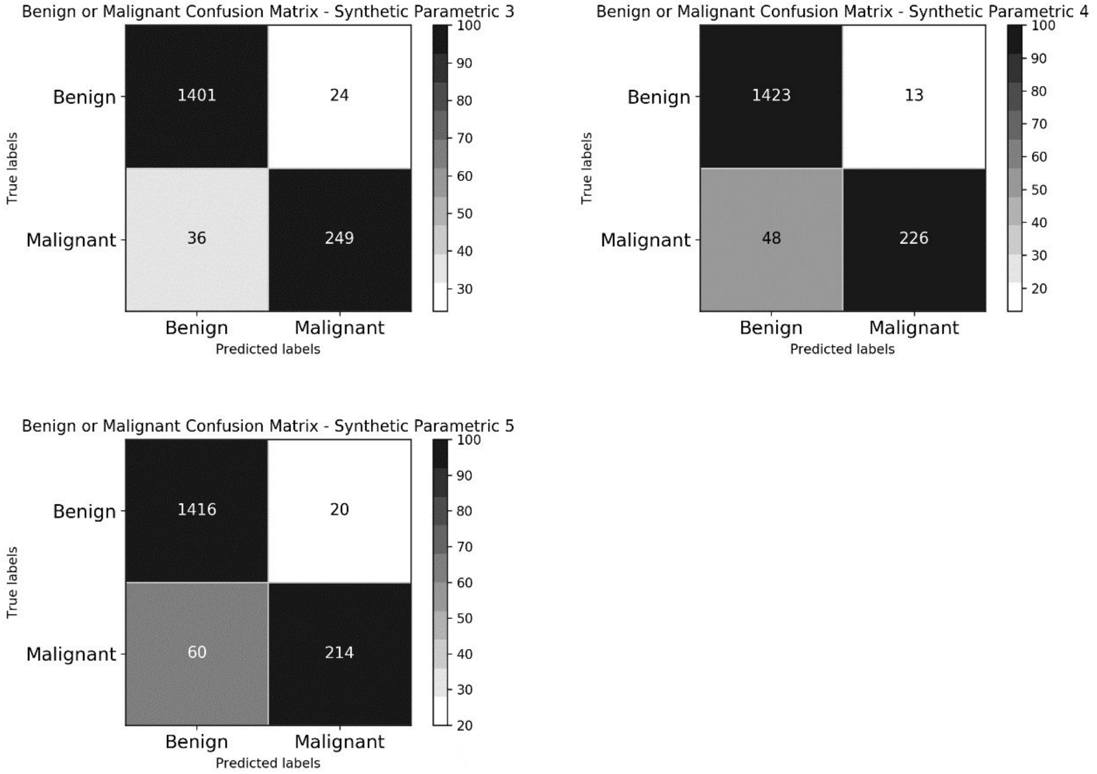
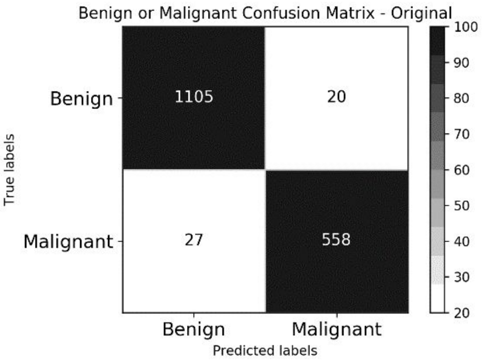
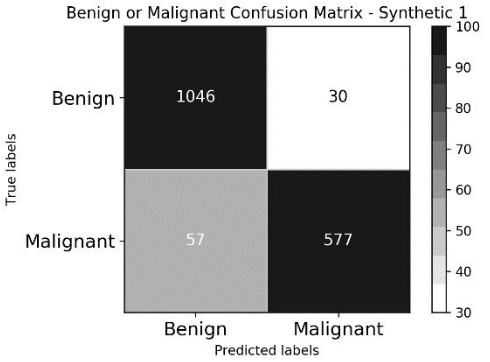
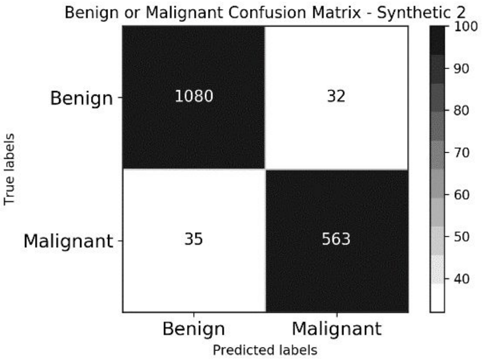
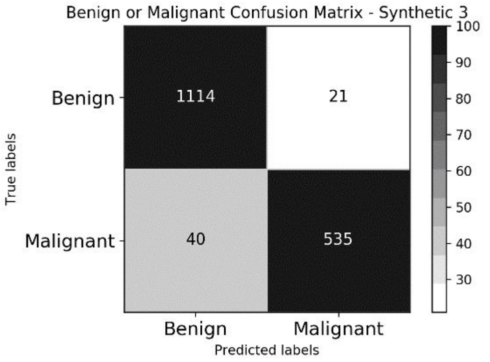


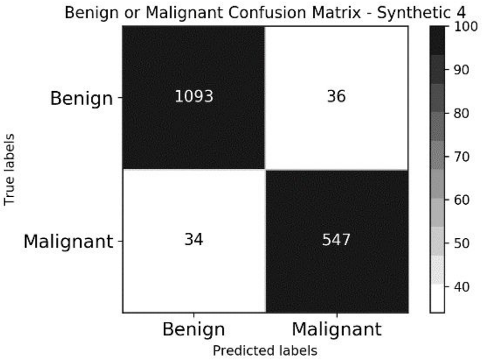
Figure 3.1.8 Confusion Matrices for the Random Forest Model when applied to each of the 11 datasets (1 original and 10 synthetic)

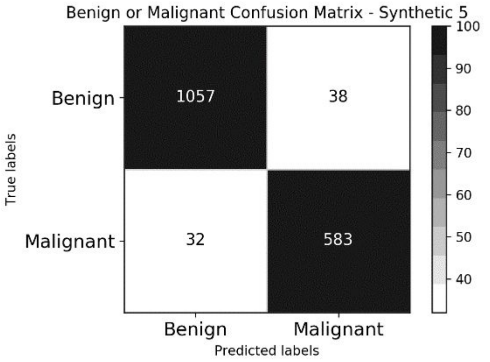

Benign or Malignant Confusion Matrix - Original
Benign
Malignant
True labels
1105
20
27
558
Benign
Malignant
Predicted labels
100
90
80
70
60
50
40
30
20

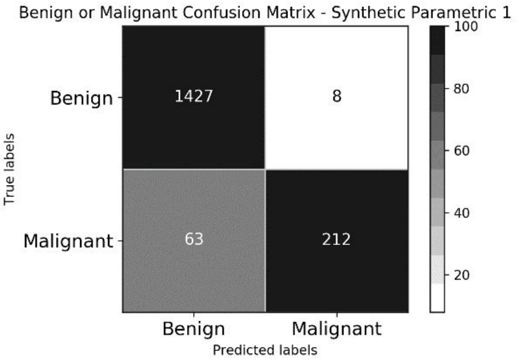

Benign or Malignant Confusion Matrix - Synthetic 1
Benign
Malignant
True labels
1046
30
57
577
Benign
Malignant
Predicted labels
100
90
80
70
60
50
40
30

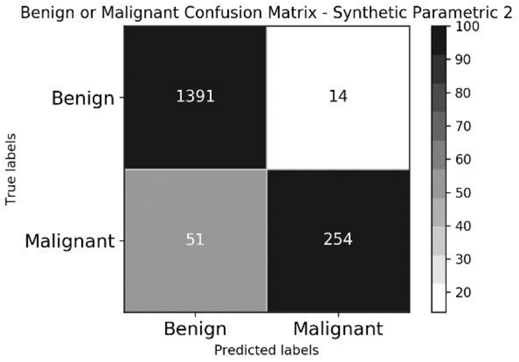

Benign or Malignant Confusion Matrix - Synthetic 2
Benign
Malignant
True labels
1080
32
35
563
Benign
Malignant
Predicted labels
100
90
80
70
60
50
40


Benign or Malignant Confusion Matrix - Synthetic 3
Benign
Malignant
True labels
1114
21
40
535
Benign
Malignant
Predicted labels
100
90
80
70
60
50
40
30


Benign or Malignant Confusion Matrix - Synthetic 4
Benign
Malignant
True labels
1093
36
34
547
Benign
Malignant
Predicted labels
100
90
80
70
60
50
40


Benign or Malignant Confusion Matrix - Synthetic 5
Benign
Malignant
True labels
1057
38
32
583
Benign
Malignant
Predicted labels
100
90
80
70
60
50
40


Benign or Malignant Confusion Matrix - Synthetic Parametric 1
Benign
Malignant
True labels
1427
8
63
212
Benign
Malignant
Predicted labels
100
80
60
40
20


Benign or Malignant Confusion Matrix - Synthetic Parametric 2
Benign
Malignant
True labels
1391
14
51
254
Benign
Malignant
Predicted labels
100
90
80
70
60
50
40
30
20

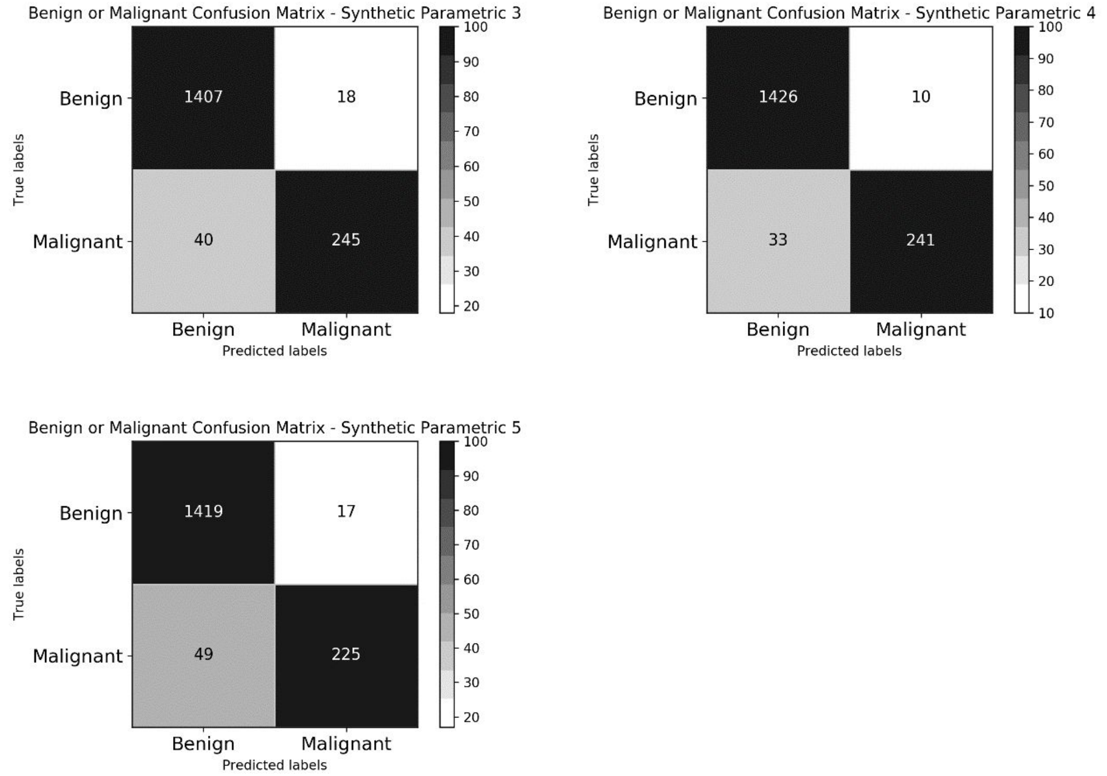


Figure 3.1.9 Confusion Matrices for the SVM Model when applied to each of the 11 datasets (1 original and 10 synthetic)

Breast Cancer Dataset Cross Comparison

A cross comparison was also carried out to determine how well classifiers that are trained on synthetic data would perform when tested with the real data. In this example the training dataset comprises 100% of the dataset listed in column 1 of Table 3.1.5 and the test set for each comprises 100% of the original dataset. Table 3.1.5 and figure 3.1.10 illustrate the accuracy scores. We observe high accuracy across all models trained on all synthetic datasets and tested on the real data. In this case, non-parametric synthetic data slightly outperforms parametric synthetic data, and the decision tree and linear models produce the lowest accuracy, whilst SVM achieves the highest average accuracy. The differences are again negligible.

Table 3.1.5 Comparison of accuracy scores achieved by each model when trained with 100% of the dataset listed in column one and tested with 100% of the original dataset.

Training Dataset (100%)	Linear Model	Decision Tree	KNN	Random Forest	SVM
Original	0.975	**1.000**	0.972	0.999	0.974
Synthetic Non-Parametric V1	0.965	0.953	**0.968**	0.965	0.966
Synthetic Non-Parametric V2	0.959	0.936	0.966	0.962	**0.971**
Synthetic Non-Parametric V3	**0.974**	0.953	0.968	0.963	0.966
Synthetic Non-Parametric V4	0.961	0.955	0.962	**0.975**	0.971
Synthetic Non-Parametric V5	0.922	0.947	0.962	0.965	**0.966**
Synthetic Parametric V1	0.874	0.895	0.924	0.912	**0.927**
Synthetic Parametric V2	**0.949**	0.909	0.936	0.931	0.939
Synthetic Parametric V3	0.842	0.900	0.928	0.930	**0.941**
Synthetic Parametric V4	**0.971**	0.915	0.924	0.930	0.939
Synthetic Parametric V5	**0.962**	0.895	0.936	0.931	0.944

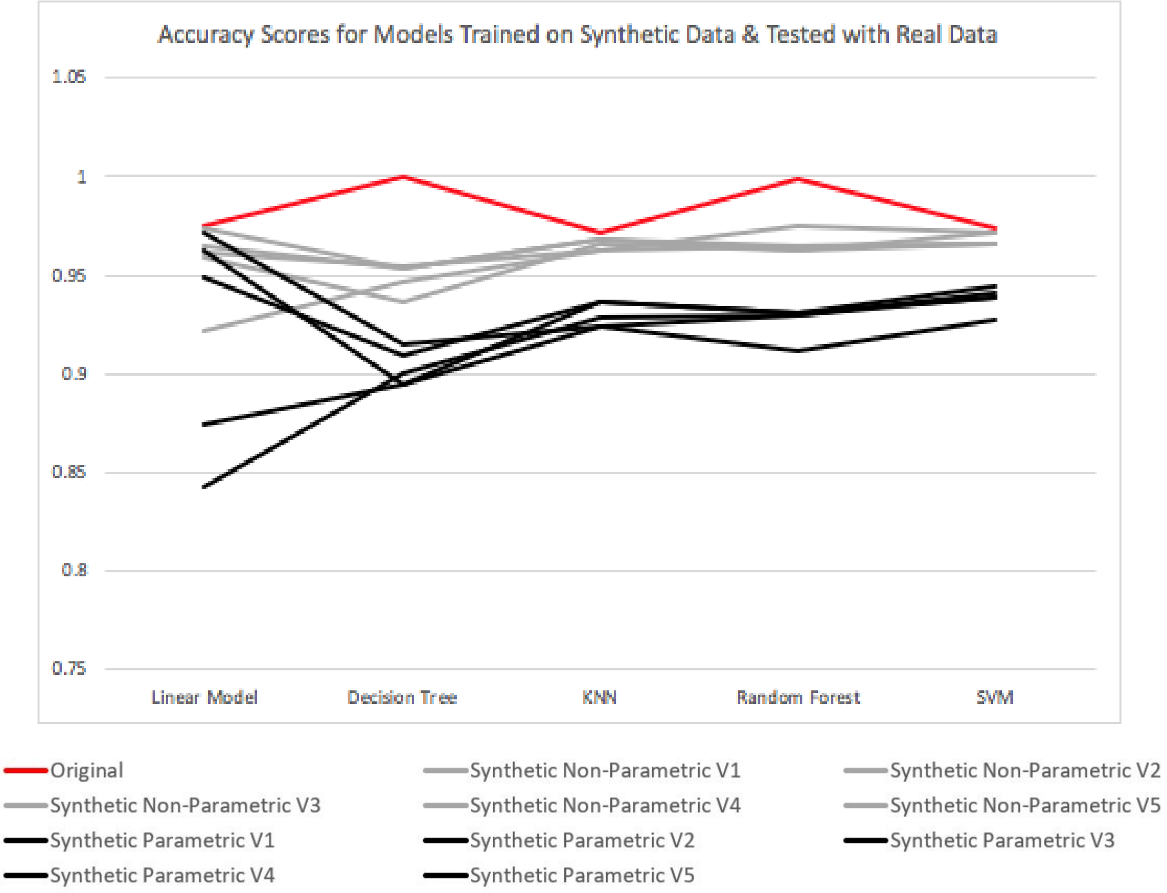


Figure 3.1.10 Comparison of accuracy scores achieved by each model when trained with 100% of the dataset listed in column one and tested with 100% of the original dataset.

Tables 3.1.6-3.1.8 illustrate the precision, recall and F1 measures, respectively for each of the five classification models after being trained by each dataset and tested with the original dataset. In cross comparisons, precision, recall and F1 scores reflect the high accuracy scores across each model. Visualisations of the decision trees trained from this data are provided in Appendix B for reference.

Table 3.1.6 Comparison of precision scores achieved by each model when trained with 100% of the dataset listed in column one and tested with 100% of the original dataset.

Training Dataset (100%)	Linear Model	Decision Tree	KNN	Random Forest	SVM
Original	0.971	1.000	0.971	0.999	0.999
Synthetic Non-Parametric V1	0.961	0.946	0.965	0.958	0.958
Synthetic Non-Parametric V2	0.948	0.929	0.964	0.961	0.961
Synthetic Non-Parametric V3	0.966	0.948	0.966	0.958	0.958
Synthetic Non-Parametric V4	0.963	0.951	0.960	0.971	0.971
Synthetic Non-Parametric V5	0.938	0.949	0.963	0.960	0.960

Synthetic Parametric V1	0.914	0.911	0.942	0.933	0.933
Synthetic Parametric V2	0.954	0.921	0.949	0.946	0.946
Synthetic Parametric V3	0.899	0.920	0.945	0.946	0.946
Synthetic Parametric V4	0.968	0.933	0.944	0.946	0.946
Synthetic Parametric V5	0.963	0.920	0.949	0.946	0.946

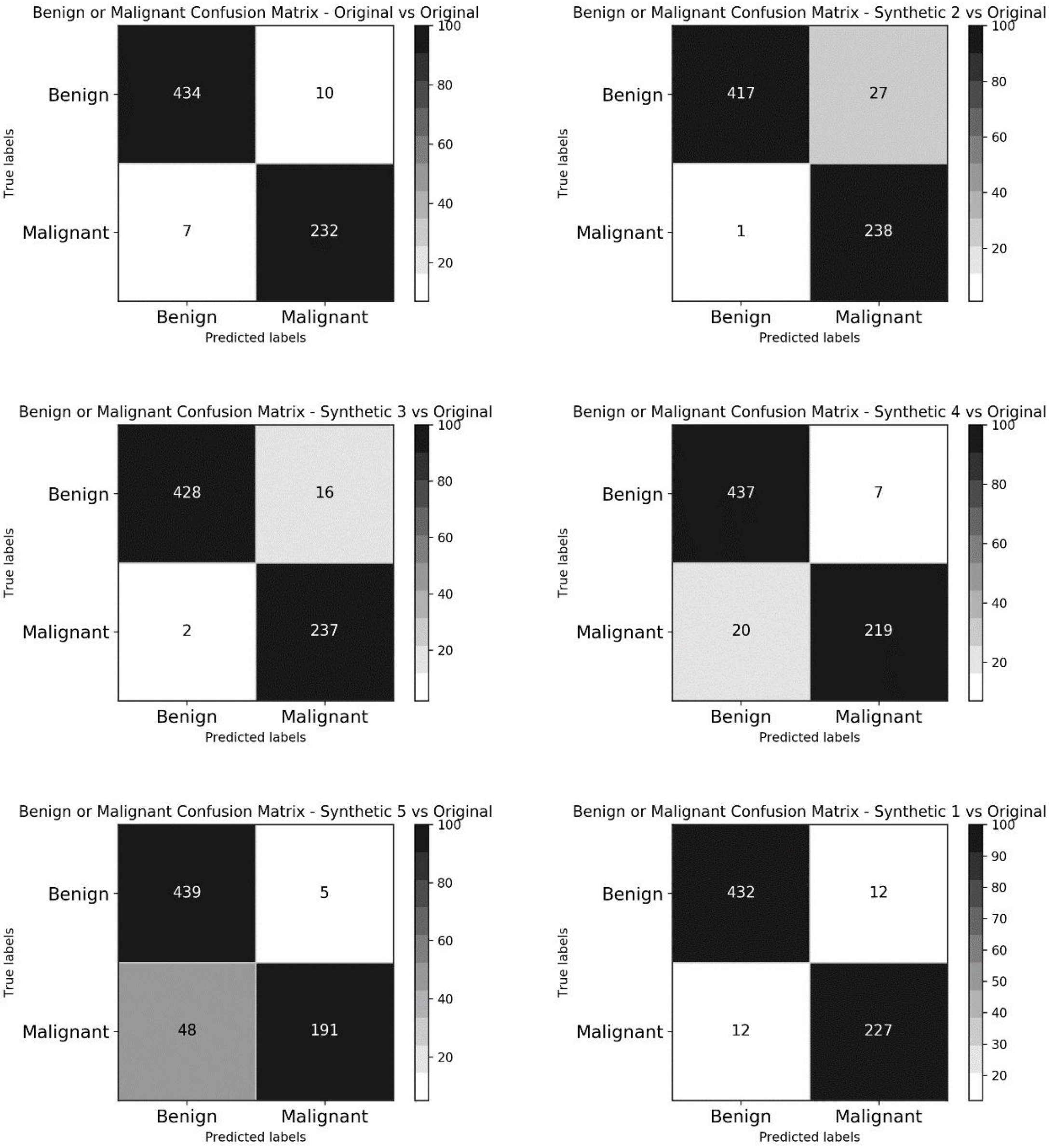
Table 3.1.7 Comparison of recall scores achieved by each model when trained with 100% of the dataset listed in column one and tested with 100% of the original dataset.

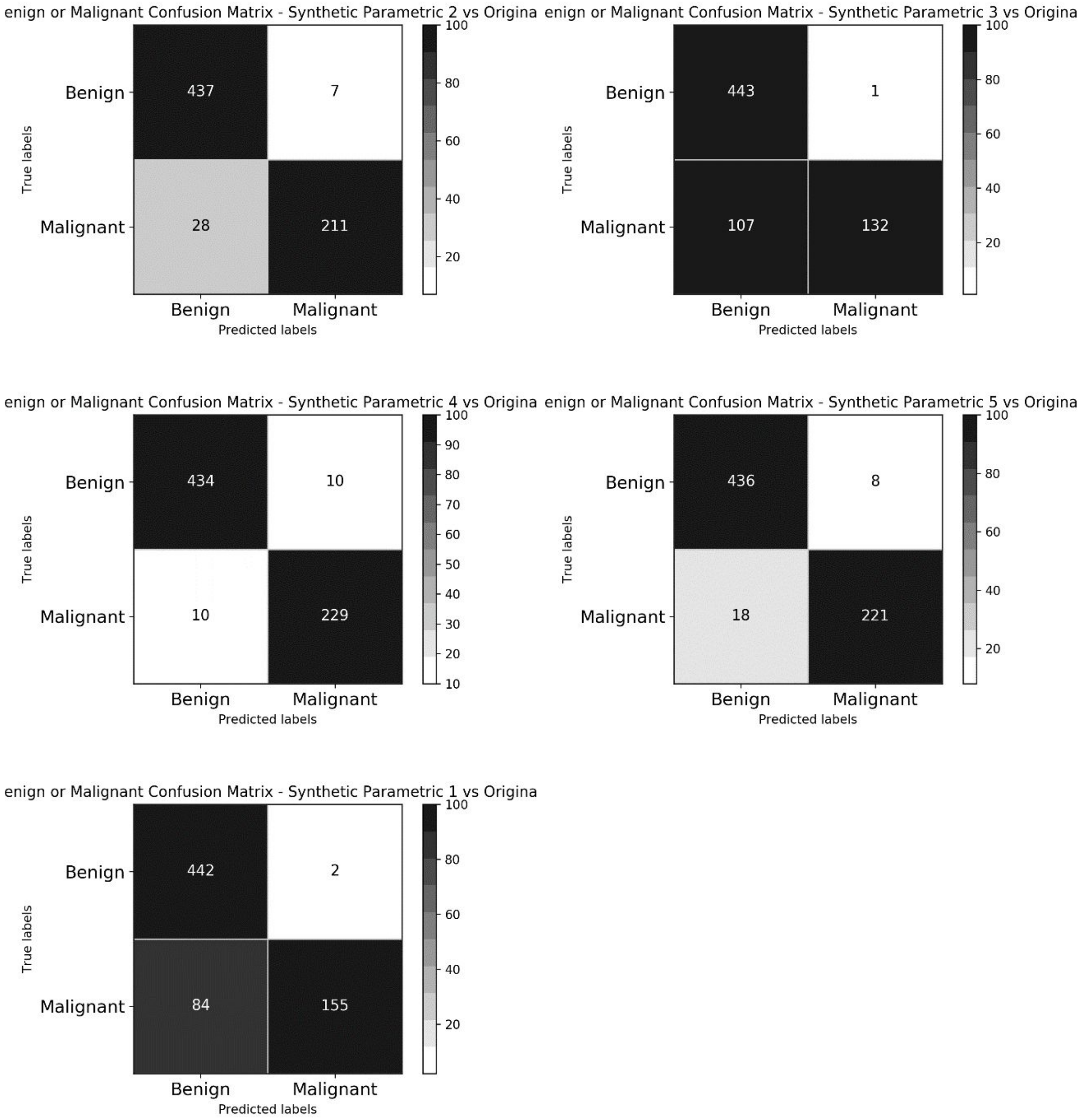
Training Dataset (100%)	Linear Model	Decision Tree	KNN	Random Forest	SVM
Original	0.974	1.000	0.968	0.998	0.998
Synthetic Non-Parametric V1	0.961	0.952	0.965	0.966	0.966
Synthetic Non-Parametric V2	0.968	0.929	0.962	0.955	0.955
Synthetic Non-Parametric V3	0.978	0.950	0.963	0.962	0.962
Synthetic Non-Parametric V4	0.950	0.949	0.956	0.975	0.975
Synthetic Non-Parametric V5	0.894	0.934	0.953	0.963	0.963
Synthetic Parametric V1	0.822	0.859	0.894	0.878	0.878
Synthetic Parametric V2	0.934	0.880	0.912	0.906	0.906
Synthetic Parametric V3	0.775	0.865	0.900	0.903	0.903
Synthetic Parametric V4	0.968	0.884	0.893	0.903	0.903
Synthetic Parametric V5	0.953	0.854	0.912	0.906	0.906

Table 3.1.8 Comparison of f1 scores achieved by each model when trained with 100% of the dataset listed in column one and tested with 100% of the original dataset.

Training Dataset (100%)	Linear Model	Decision Tree	KNN	Random Forest	SVM
Original	0.973	1.000	0.969	0.998	0.998
Synthetic Non-Parametric V1	0.961	0.949	0.965	0.962	0.962
Synthetic Non-Parametric V2	0.956	0.929	0.963	0.958	0.958
Synthetic Non-Parametric V3	0.971	0.949	0.965	0.960	0.960
Synthetic Non-Parametric V4	0.956	0.950	0.958	0.973	0.973
Synthetic Non-Parametric V5	0.911	0.941	0.958	0.962	0.962
Synthetic Parametric V1	0.847	0.877	0.912	0.898	0.898
Synthetic Parametric V2	0.943	0.896	0.926	0.921	0.921
Synthetic Parametric V3	0.801	0.884	0.917	0.919	0.919
Synthetic Parametric V4	0.968	0.902	0.912	0.919	0.919
Synthetic Parametric V5	0.958	0.876	0.926	0.921	0.921

The confusion matrices for the performance of each of the five classifiers, trained on 100% of each of the eleven datasets (original, 5 synthetic non-parametric and 5 synthetic parametric) and tested on 100% of the original dataset are shown in Figure 3.1.11-3.1.15 for the Linear model, Decision Tree model, KNN model, Random Forest model and SVM model respectively. We again observe that the majority of observations are classified correctly with only small instances of false positives and false negatives present for all models trained using synthetic data and tested using the real data. Therefore the synthetic data for this dataset is shown to be suitable for training classification models that can then adequately classify new, real records.





enign or Malignant Confusion Matrix - Synthetic Parametric 2 vs Origina

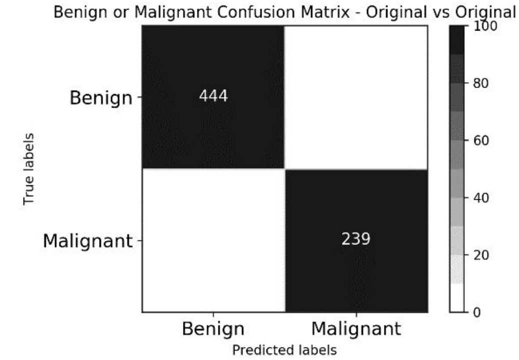
enign or Malignant Confusion Matrix - Synthetic Parametric 3 vs Origina

enign or Malignant Confusion Matrix - Synthetic Parametric 4 vs Origina

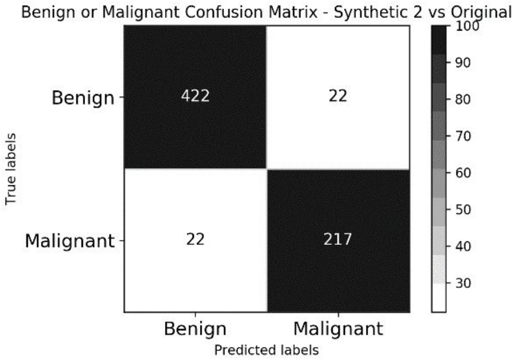
enign or Malignant Confusion Matrix - Synthetic Parametric 5 vs Origina

enign or Malignant Confusion Matrix - Synthetic Parametric 1 vs Origina

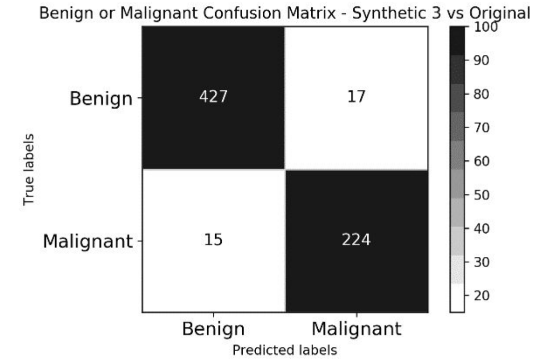
Figure 3.1.11 Confusion Matrices for the Linear Model when trained with each of the 11 datasets (1 original and 10 synthetic) and tested on 100% of the original dataset



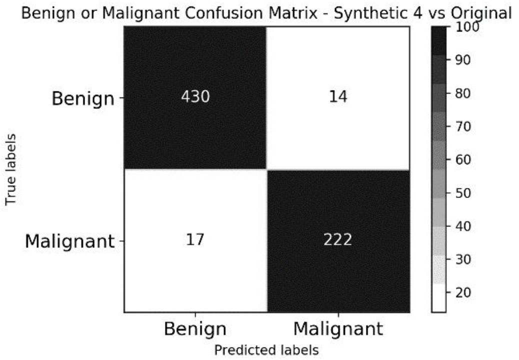
Benign or Malignant Confusion Matrix - Original vs Original



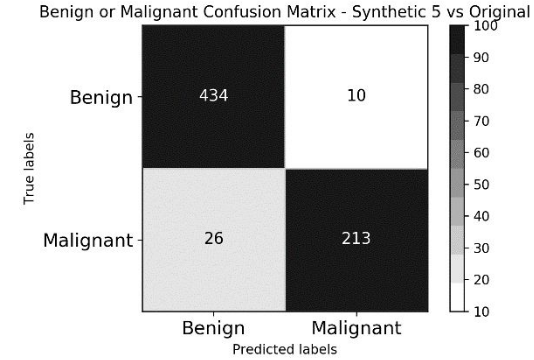
Benign or Malignant Confusion Matrix - Synthetic 2 vs Original



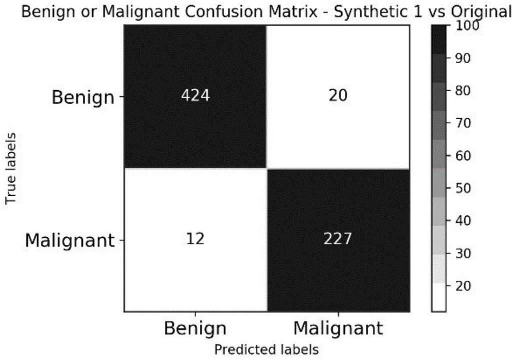
Benign or Malignant Confusion Matrix - Synthetic 3 vs Original



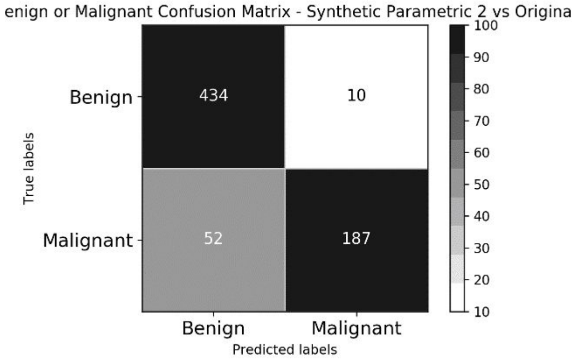
Benign or Malignant Confusion Matrix - Synthetic 4 vs Original



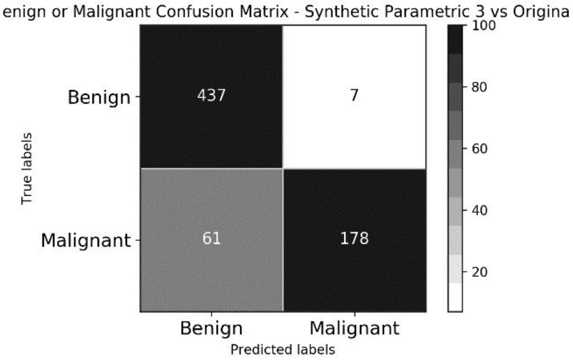
Benign or Malignant Confusion Matrix - Synthetic 5 vs Original



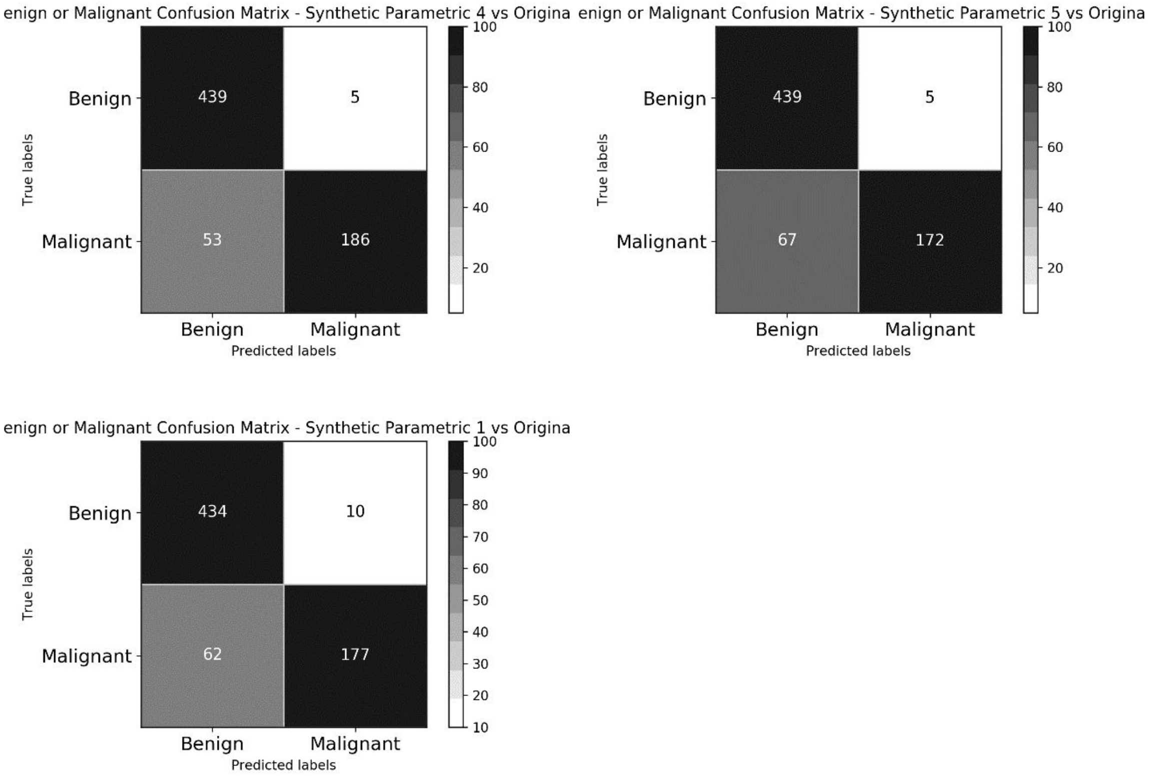
Benign or Malignant Confusion Matrix - Synthetic 1 vs Original



enign or Malignant Confusion Matrix - Synthetic Parametric 2 vs Origina



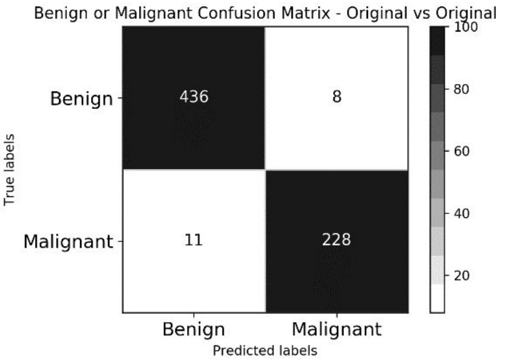
enign or Malignant Confusion Matrix - Synthetic Parametric 3 vs Origina

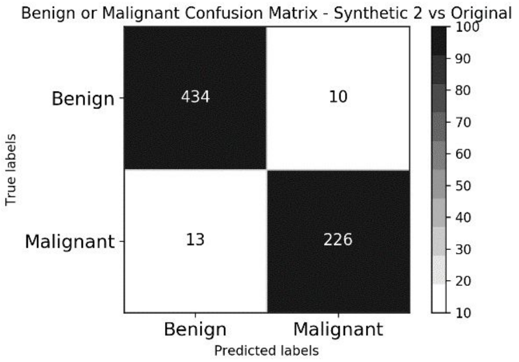

enign or Malignant Confusion Matrix - Synthetic Parametric 4 vs Origina

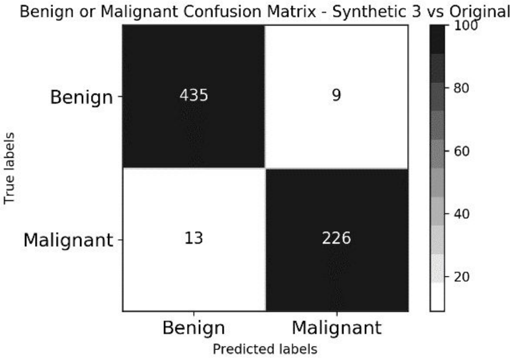
enign or Malignant Confusion Matrix - Synthetic Parametric 5 vs Origina

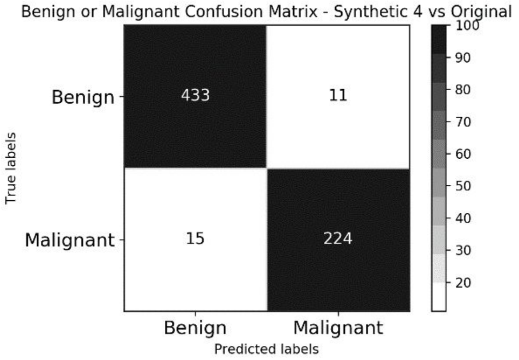
enign or Malignant Confusion Matrix - Synthetic Parametric 1 vs Origina

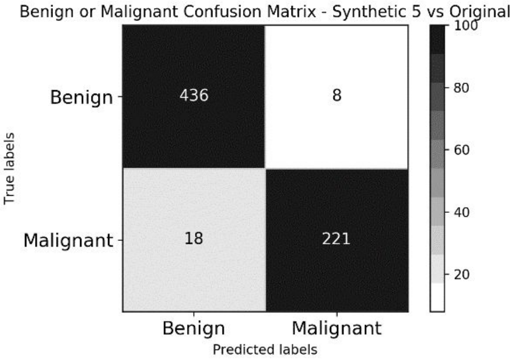
Figure 3.1.12 Confusion Matrices for the Decision Tree Model when trained with each of the 11 datasets (1 original and 10 synthetic) and tested on 100% of the original dataset

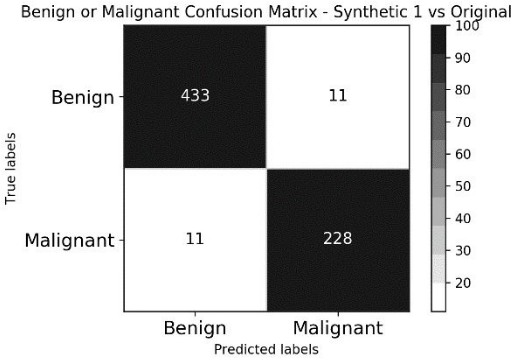

Benign or Malignant Confusion Matrix - Original vs Original
True labels
Benign
Malignant
436
8
11
228
Benign
Malignant
Predicted labels
100
80
60
40
20

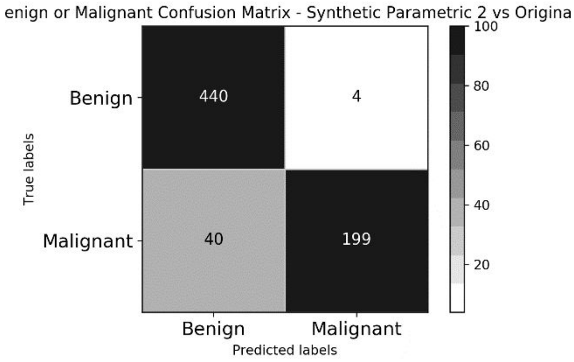

Benign or Malignant Confusion Matrix - Synthetic 2 vs Original
True labels
Benign
Malignant
434
10
13
226
Benign
Malignant
Predicted labels
100
90
80
70
60
50
40
30
20
10

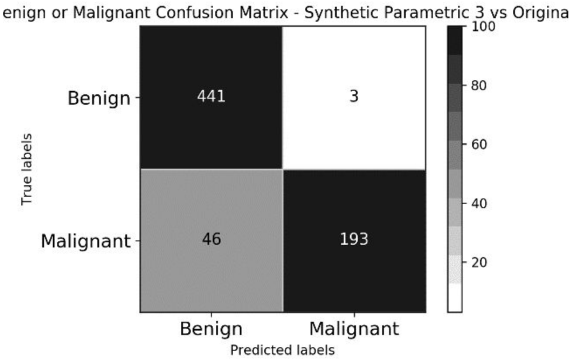

Benign or Malignant Confusion Matrix - Synthetic 3 vs Original
True labels
Benign
Malignant
435
9
13
226
Benign
Malignant
Predicted labels
100
80
60
40
20


Benign or Malignant Confusion Matrix - Synthetic 4 vs Original
True labels
Benign
Malignant
433
11
15
224
Benign
Malignant
Predicted labels
100
90
80
70
60
50
40
30
20


Benign or Malignant Confusion Matrix - Synthetic 5 vs Original
True labels
Benign
Malignant
436
8
18
221
Benign
Malignant
Predicted labels
100
80
60
40
20


Benign or Malignant Confusion Matrix - Synthetic 1 vs Original
True labels
Benign
Malignant
433
11
11
228
Benign
Malignant
Predicted labels
100
90
80
70
60
50
40
30
20


enign or Malignant Confusion Matrix - Synthetic Parametric 2 vs Origina
True labels
Benign
Malignant
440
4
40
199
Benign
Malignant
Predicted labels
100
80
60
40
20


enign or Malignant Confusion Matrix - Synthetic Parametric 3 vs Origina
True labels
Benign
Malignant
441
3
46
193
Benign
Malignant
Predicted labels
100
80
60
40
20

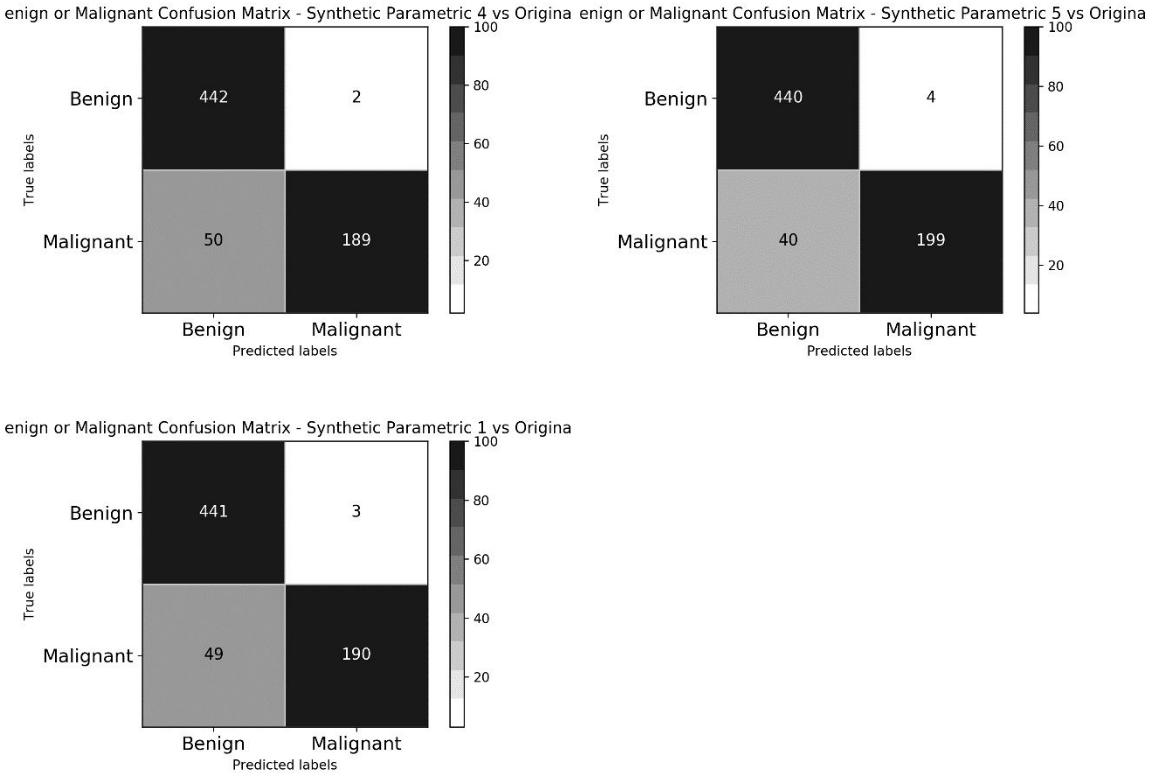
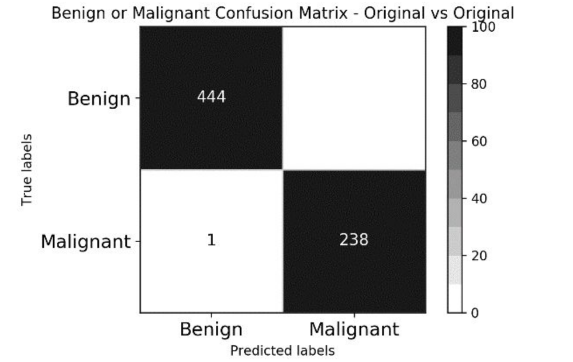
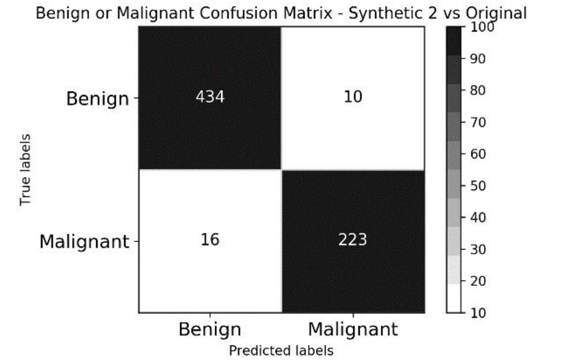
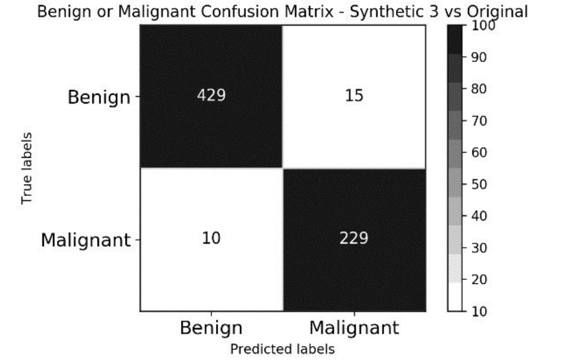
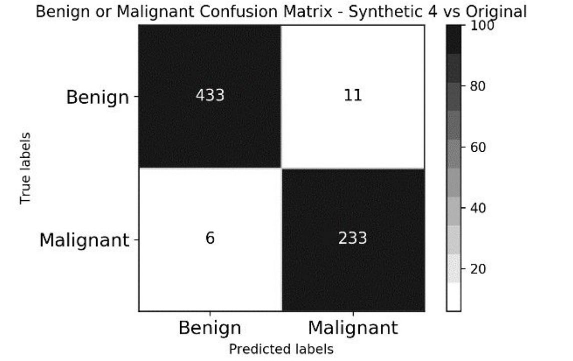


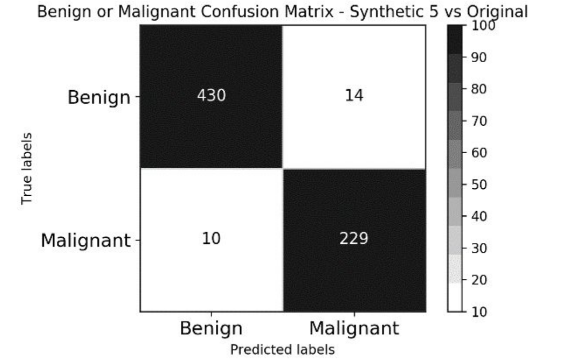
Figure 3.1.13 Confusion Matrices for the KNN Model when trained with each of the 11 datasets (1 original and 10 synthetic) and tested on 100% of the original dataset

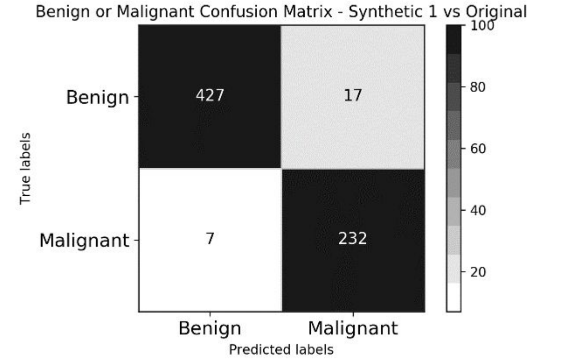

Benign or Malignant Confusion Matrix - Original vs Original
True labels
Benign
Malignant
444
1
238
Predicted labels
Benign
Malignant
100
80
60
40
20
0

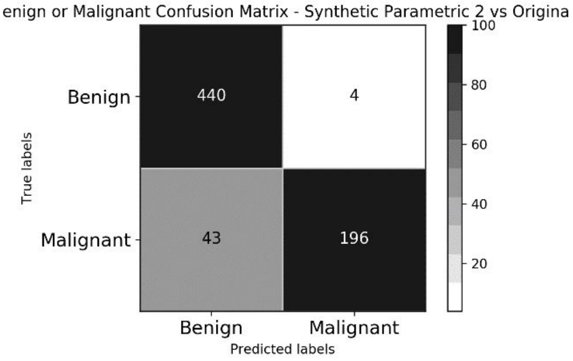

Benign or Malignant Confusion Matrix - Synthetic 2 vs Original
True labels
Benign
Malignant
434
10
16
223
Predicted labels
Benign
Malignant
100
90
80
70
60
50
40
30
20
10

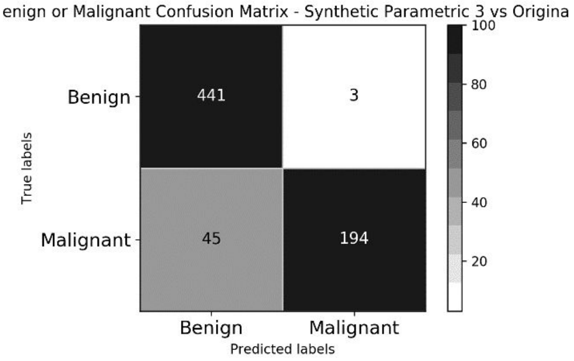

Benign or Malignant Confusion Matrix - Synthetic 3 vs Original
True labels
Benign
Malignant
429
15
10
229
Predicted labels
Benign
Malignant
100
90
80
70
60
50
40
30
20
10


Benign or Malignant Confusion Matrix - Synthetic 4 vs Original
True labels
Benign
Malignant
433
11
6
233
Predicted labels
Benign
Malignant
100
80
60
40
20


Benign or Malignant Confusion Matrix - Synthetic 5 vs Original
True labels
Benign
Malignant
430
14
10
229
Predicted labels
Benign
Malignant
100
90
80
70
60
50
40
30
20
10


Benign or Malignant Confusion Matrix - Synthetic 1 vs Original
True labels
Benign
Malignant
427
17
7
232
Predicted labels
Benign
Malignant
100
80
60
40
20


enign or Malignant Confusion Matrix - Synthetic Parametric 2 vs Origina
True labels
Benign
Malignant
440
4
43
196
Predicted labels
Benign
Malignant
100
80
60
40
20


enign or Malignant Confusion Matrix - Synthetic Parametric 3 vs Origina
True labels
Benign
Malignant
441
3
45
194
Predicted labels
Benign
Malignant
100
80
60
40
20

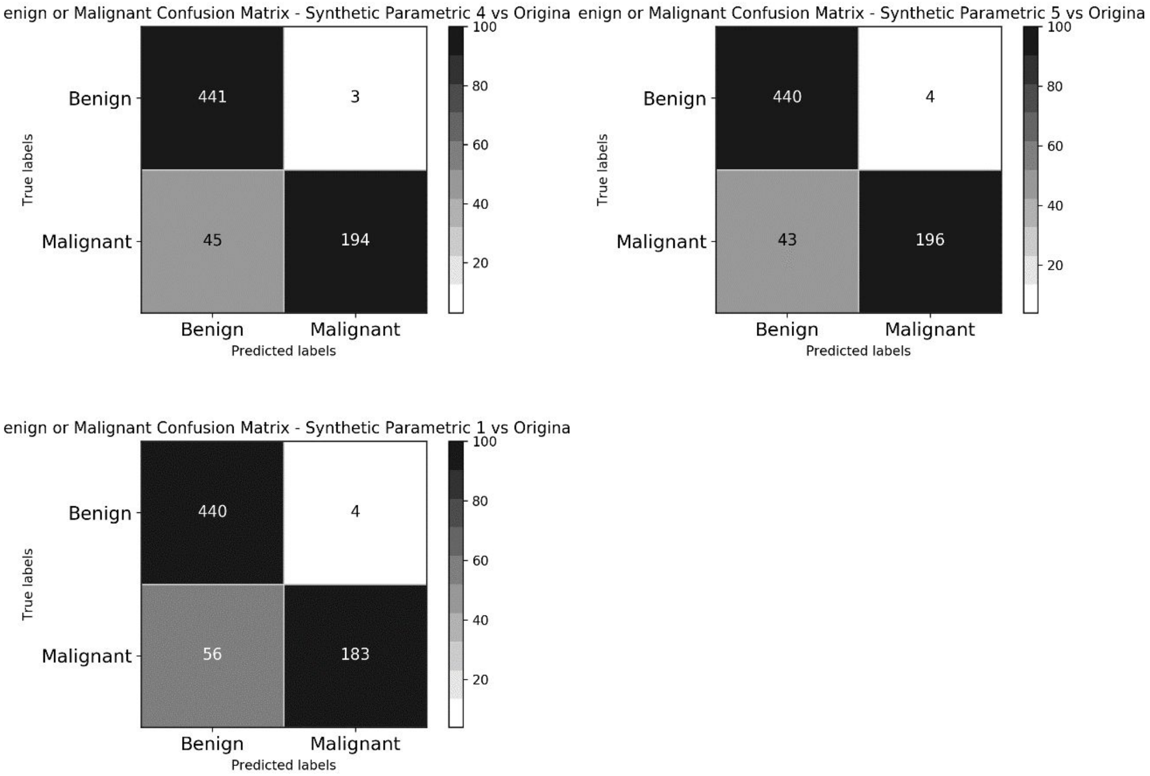
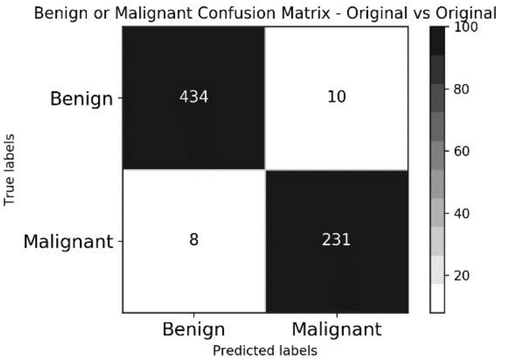
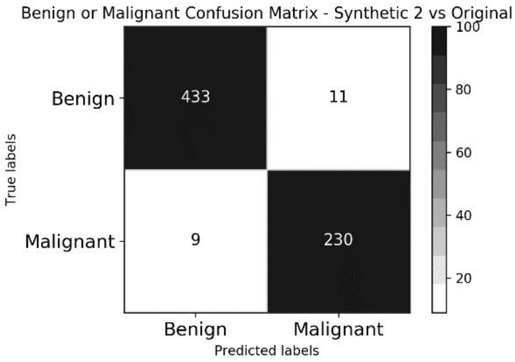
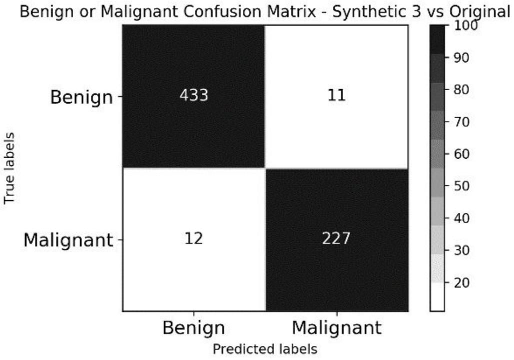
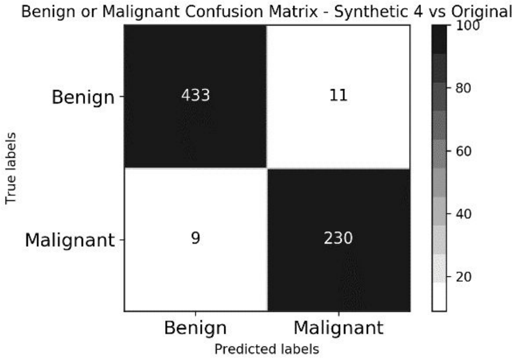


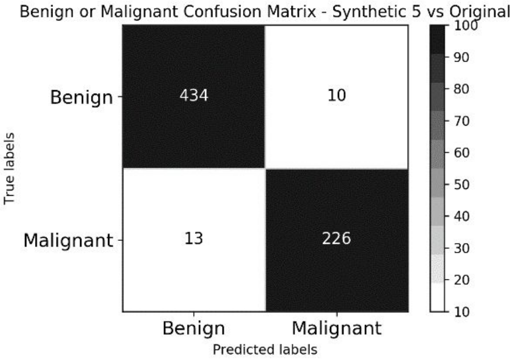
Figure 3.1.14 Confusion Matrices for the Random Forest Model when trained with each of the 11 datasets (1 original and 10 synthetic) and tested on 100% of the original dataset

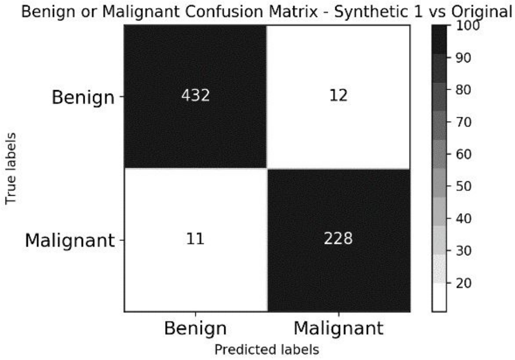

Benign or Malignant Confusion Matrix - Original vs Original
True labels
Benign
Malignant
434
10
8
231
Benign
Malignant
Predicted labels
100
80
60
40
20

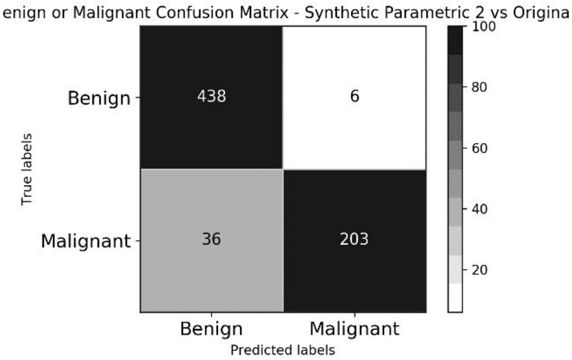

Benign or Malignant Confusion Matrix - Synthetic 2 vs Original
True labels
Benign
Malignant
433
11
9
230
Benign
Malignant
Predicted labels
100
80
60
40
20

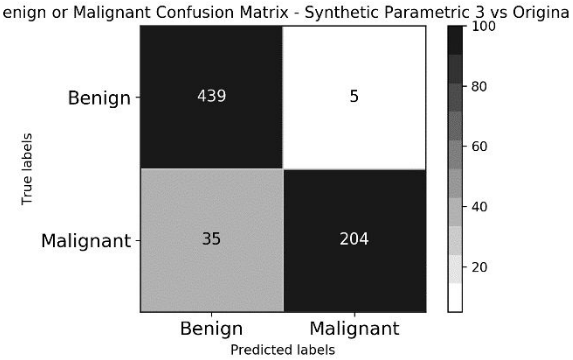

Benign or Malignant Confusion Matrix - Synthetic 3 vs Original
True labels
Benign
Malignant
433
11
12
227
Benign
Malignant
Predicted labels
100
90
80
70
60
50
40
30
20


Benign or Malignant Confusion Matrix - Synthetic 4 vs Original
True labels
Benign
Malignant
433
11
9
230
Benign
Malignant
Predicted labels
100
80
60
40
20


Benign or Malignant Confusion Matrix - Synthetic 5 vs Original
True labels
Benign
Malignant
434
10
13
226
Benign
Malignant
Predicted labels
100
90
80
70
60
50
40
30
20
10


Benign or Malignant Confusion Matrix - Synthetic 1 vs Original
True labels
Benign
Malignant
432
12
11
228
Benign
Malignant
Predicted labels
100
90
80
70
60
50
40
30
20


enign or Malignant Confusion Matrix - Synthetic Parametric 2 vs Origina
True labels
Benign
Malignant
438
6
36
203
Benign
Malignant
Predicted labels
100
80
60
40
20


enign or Malignant Confusion Matrix - Synthetic Parametric 3 vs Origina
True labels
Benign
Malignant
439
5
35
204
Benign
Malignant
Predicted labels
100
80
60
40
20

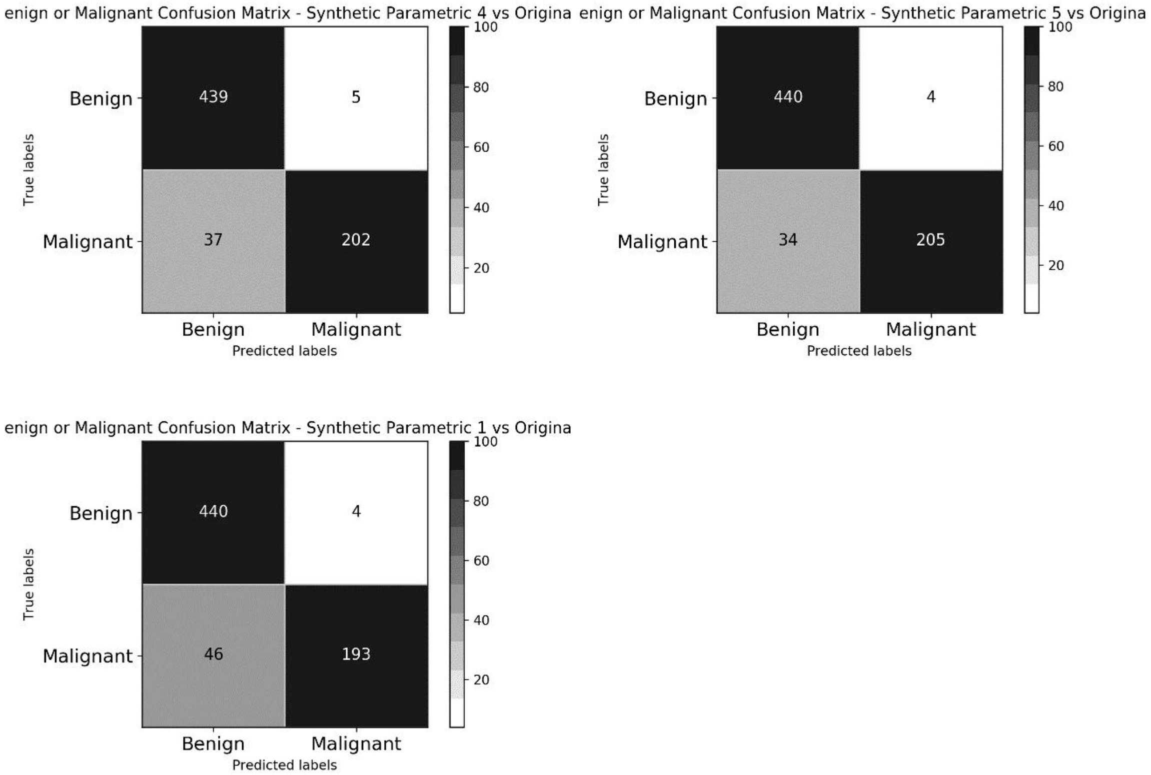


Figure 3.1.15 Confusion Matrices for the SVM Model when trained with each of the 11 datasets (1 original and 10 synthetic) and tested on 100% of the original dataset

3.2 Nursery Dataset Results

To compare the performance of each model after being trained with the original and synthetic Nursery datasets, again evaluation metrics accuracy, precision, recall and F1 score were computed and the results are shown in Tables 3.2.1-3.2.4 and Figures 3.2.1-3.2.4, respectively. These metrics are calculated for the five classification models after being trained by the original dataset and the 10 synthetic datasets (five non-parametric and five parametric). In each case, 10-fold cross-validation is utilised with a train/test split of 75/25.

Table 3.2.1 Comparison of accuracy scores achieved by each model as trained by each Nursery dataset

Dataset	Linear Model	Decision Tree	KNN	Random Forest	SVM
Original	0.897	0.969	0.964	0.960	**0.972**
Synthetic Non-Parametric V1	0.889	0.961	0.948	0.959	**0.964**
Synthetic Non-Parametric V2	0.884	0.964	0.952	0.960	**0.965**
Synthetic Non-Parametric V3	0.895	0.959	0.949	0.961	**0.964**
Synthetic Non-Parametric V4	0.885	**0.964**	0.950	0.959	**0.964**
Synthetic Non-Parametric V5	0.891	0.964	0.950	0.961	**0.965**
Synthetic Parametric V1	0.889	0.899	0.893	0.902	**0.918**
Synthetic Parametric V2	0.896	0.902	0.895	0.907	**0.921**
Synthetic Parametric V3	0.885	0.894	0.892	0.905	**0.917**
Synthetic Parametric V4	0.896	0.902	0.895	0.907	**0.917**
Synthetic Parametric V5	0.892	0.903	0.898	0.907	**0.922**

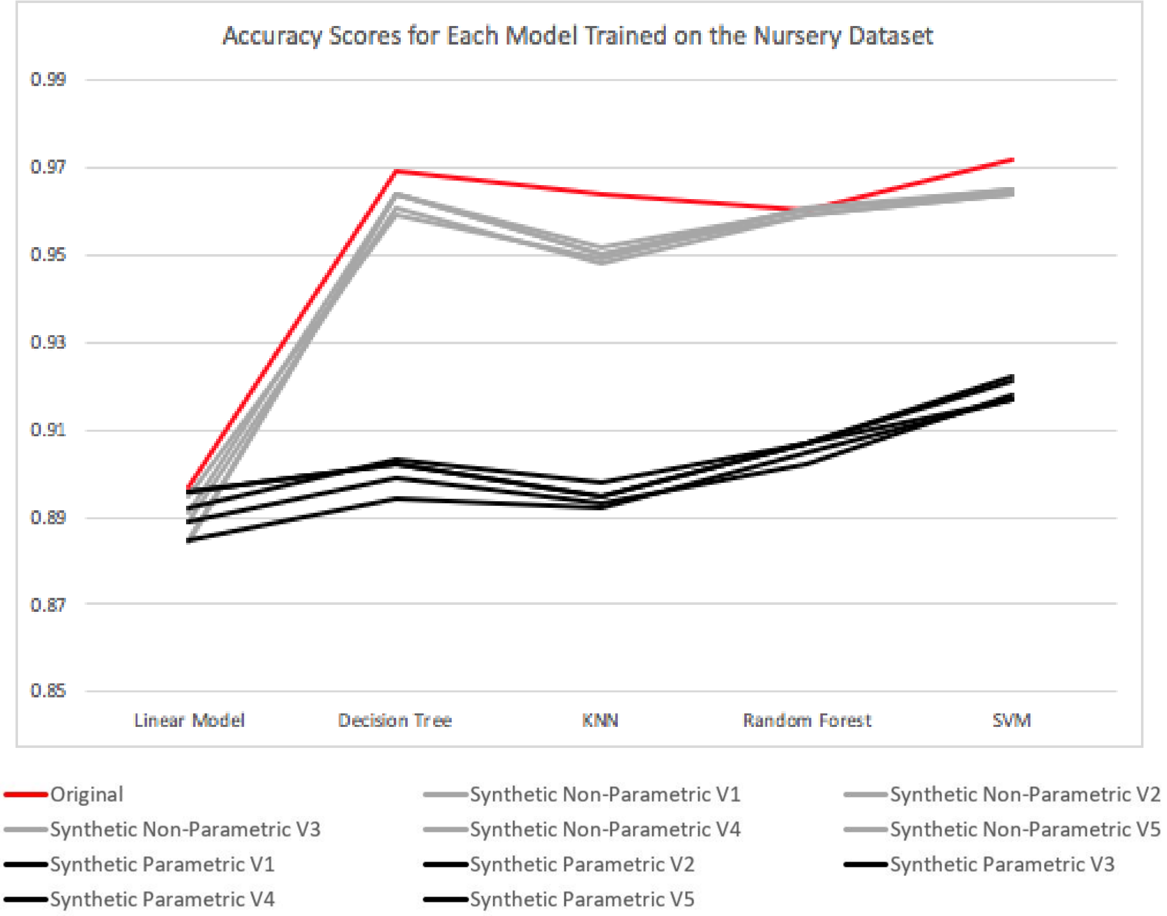


Figure 3.2.1 Comparison of accuracy scores achieved by each model as trained by each Nursery dataset

We observe that all models perform well on the original and synthetic datasets with a minimum, yet still high, accuracy above 0.88. Models trained on the non-parametric synthetic data produce more favourable results than those trained on the parametric synthetic datasets. The performance of the models on the non-parametric synthetic data compared with the real data demonstrates very minor differences, whereas parametric data does not perform to the same degree.

Tables 3.2.2-3.2.4 and figures 3.2.2-3.2.4 illustrate the precision, recall and F1 measures, respectively, for each of the five classification models after being trained by the original dataset and the ten synthetic datasets. We observe similar trends in these metrics as were observed for accuracy, however precision, recall and F1 scores are lower overall.

Table 3.2.2 Comparison of precision scores achieved by each model as trained by each dataset

Dataset	Linear Model	Decision Tree	KNN	Random Forest	SVM
Original	0.755	0.887	0.913	0.904	**0.918**
Synthetic Non-Parametric V1	0.667	0.755	0.778	0.786	**0.791**
Synthetic Non-Parametric V2	0.689	0.799	0.822	0.828	**0.833**
Synthetic Non-Parametric V3	0.848	0.913	0.932	0.948	**0.952**
Synthetic Non-Parametric V4	0.723	0.833	0.836	0.849	**0.851**
Synthetic Non-Parametric V5	0.788	0.924	0.951	0.968	**0.973**
Synthetic Parametric V1	0.691	0.761	0.770	0.784	**0.815**
Synthetic Parametric V2	0.642	0.709	0.719	0.741	**0.765**
Synthetic Parametric V3	0.652	0.748	0.785	0.790	**0.827**
Synthetic Parametric V4	0.653	0.714	0.736	0.737	**0.779**
Synthetic Parametric V5	0.691	0.710	0.735	0.739	**0.767**

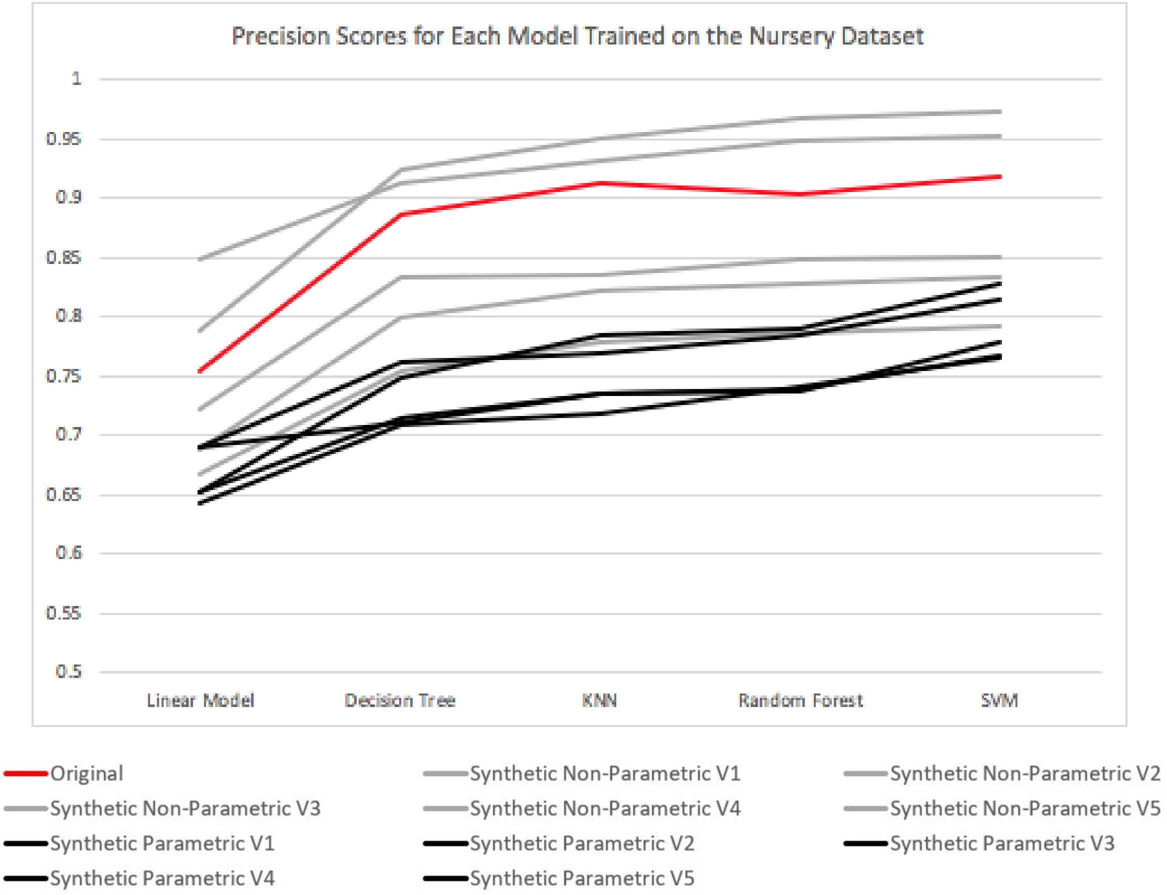


Figure 3.2.2 Comparison of precision scores achieved by each model as trained by each dataset

Table 3.2.3 Comparison of recall scores achieved by each model as trained by each dataset

Dataset	Linear Model	Decision Tree	KNN	Random Forest	SVM
Original	0.658	**0.859**	0.804	0.800	0.799
Synthetic Non-Parametric V1	0.582	**0.753**	0.705	0.721	0.715
Synthetic Non-Parametric V2	0.602	**0.796**	0.742	0.753	0.744
Synthetic Non-Parametric V3	0.698	**0.870**	0.831	0.845	0.824
Synthetic Non-Parametric V4	0.619	**0.805**	0.753	0.771	0.766
Synthetic Non-Parametric V5	0.702	**0.906**	0.868	0.873	0.867
Synthetic Parametric V1	0.612	**0.691**	0.669	0.667	0.684
Synthetic Parametric V2	0.583	**0.678**	0.626	0.638	0.655
Synthetic Parametric V3	0.623	**0.732**	0.689	0.690	0.690
Synthetic Parametric V4	0.585	**0.695**	0.635	0.646	0.633
Synthetic Parametric V5	0.594	**0.701**	0.659	0.661	0.675

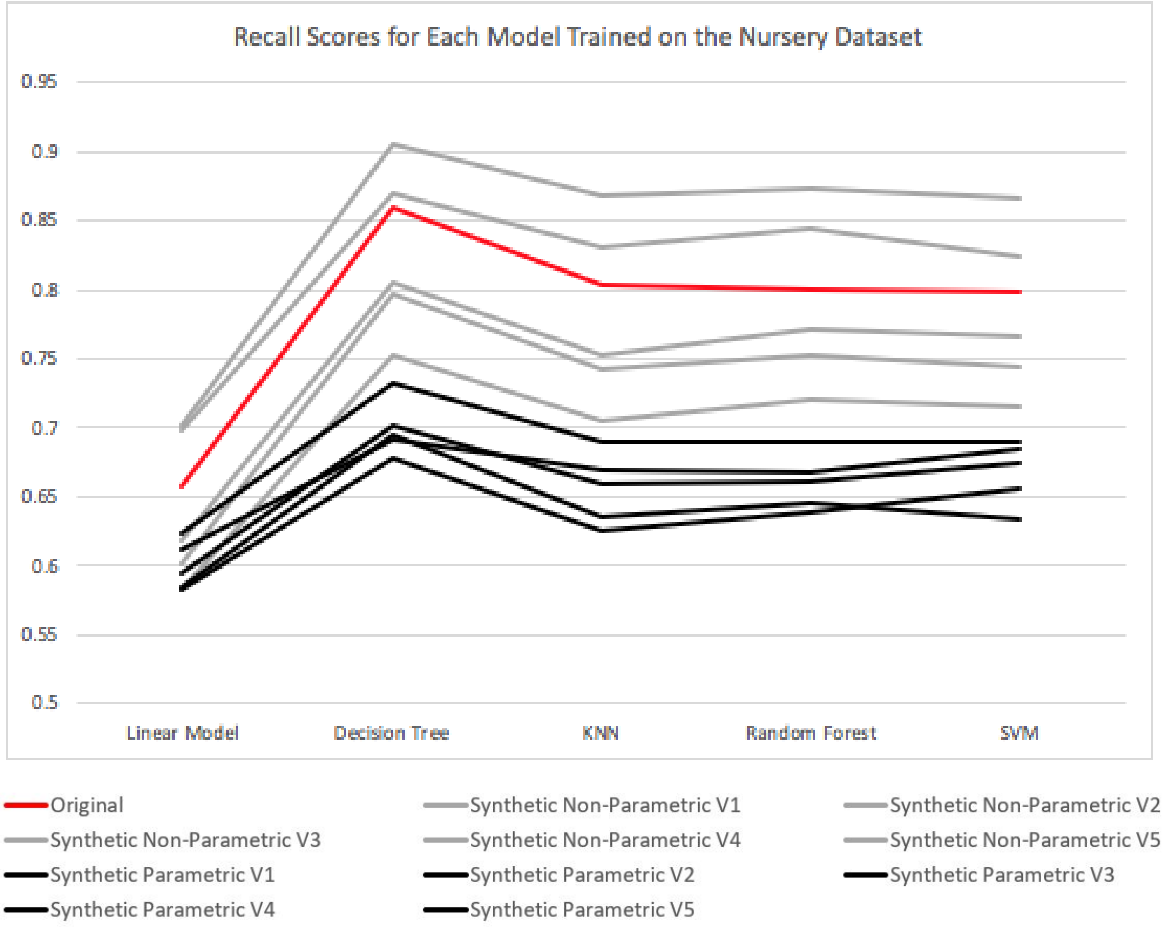


Figure 3.2.3 Comparison of recall scores achieved by each model as trained by each dataset

Table 3.2.4 Comparison of f1 scores achieved by each model as trained by each dataset

Dataset	Linear Model	Decision Tree	KNN	Random Forest	SVM
Original	0.658	**0.871**	0.838	0.832	0.832
Synthetic Non-Parametric V1	0.585	**0.752**	0.731	0.745	0.741
Synthetic Non-Parametric V2	0.607	**0.796**	0.770	0.780	0.774
Synthetic Non-Parametric V3	0.703	**0.887**	0.864	0.879	0.860
Synthetic Non-Parametric V4	0.622	**0.818**	0.782	0.799	0.796
Synthetic Non-Parametric V5	0.705	**0.914**	0.899	0.907	0.902
Synthetic Parametric V1	0.613	**0.715**	0.695	0.692	0.713
Synthetic Parametric V2	0.584	**0.690**	0.650	0.664	0.684
Synthetic Parametric V3	0.620	**0.739**	0.715	0.715	0.719
Synthetic Parametric V4	0.584	**0.703**	0.660	0.670	0.655
Synthetic Parametric V5	0.596	**0.705**	0.684	0.685	0.703

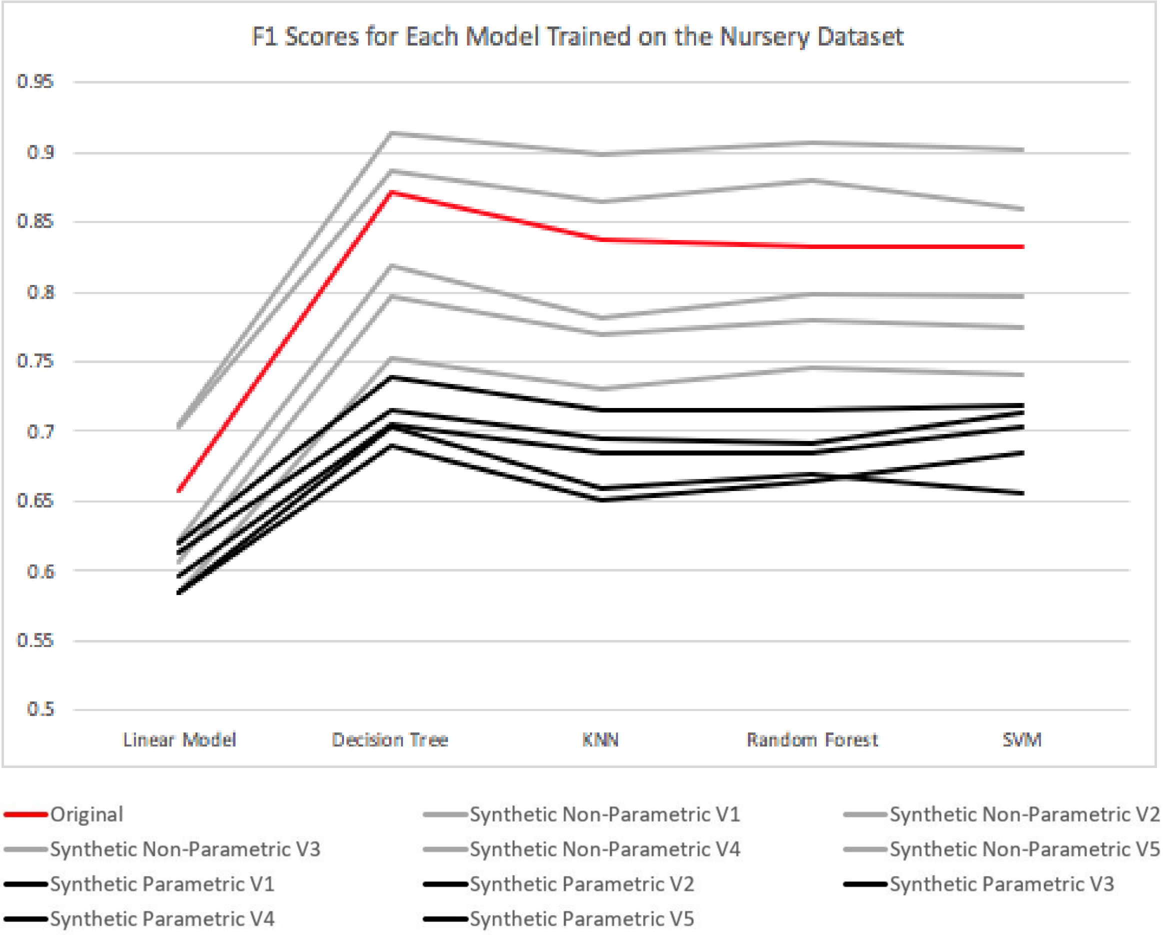
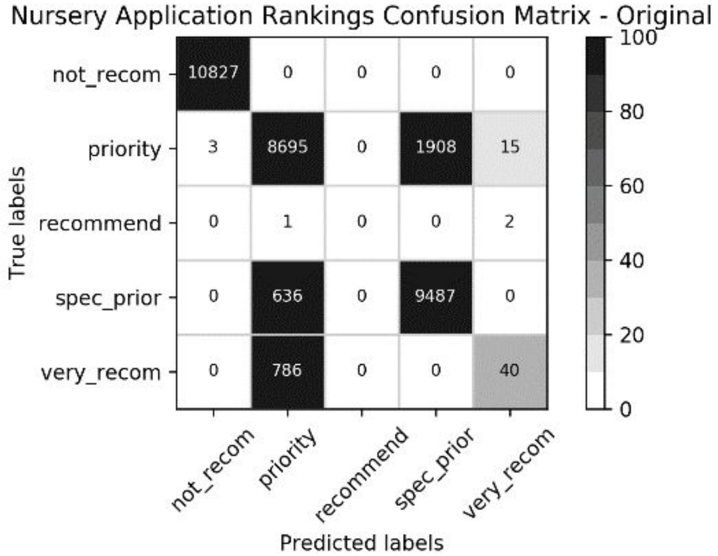
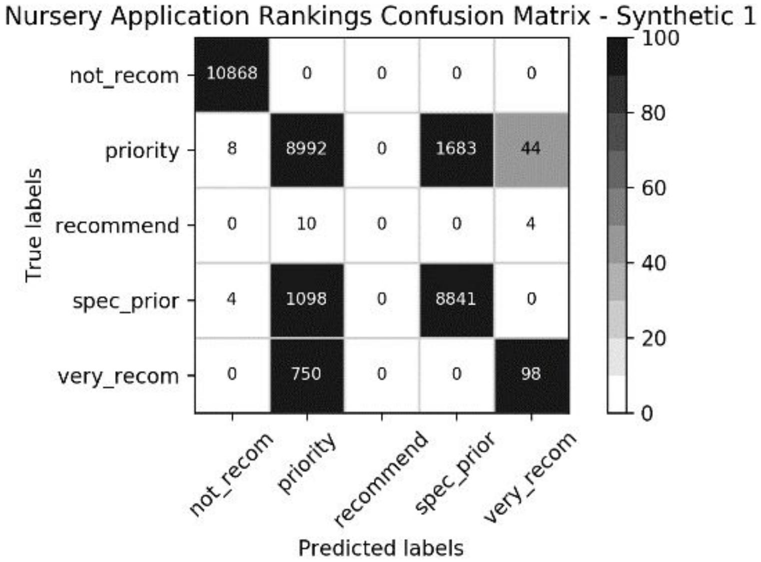
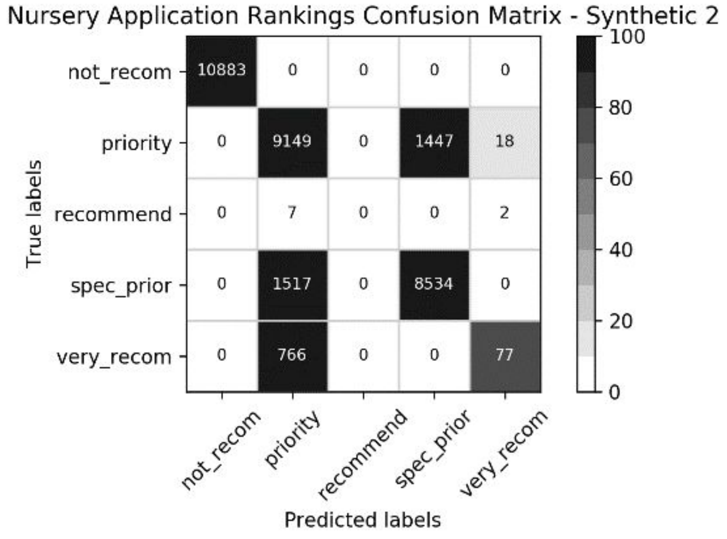


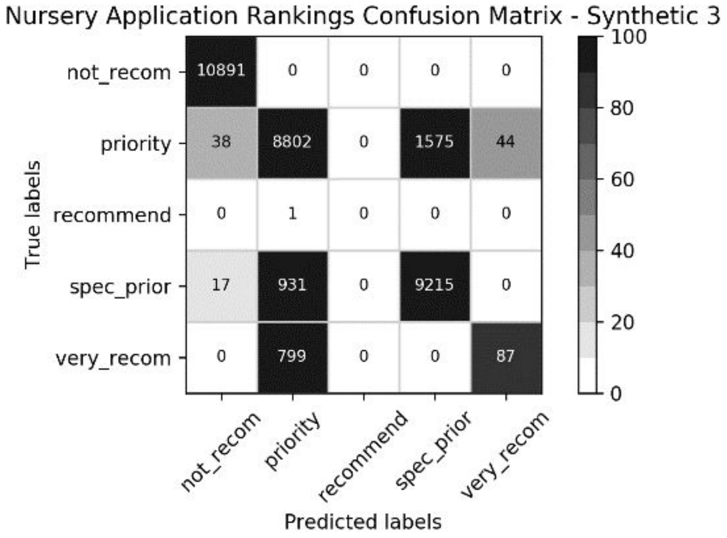
Figure 3.2.4 Comparison of f1 scores achieved by each model as trained by each dataset

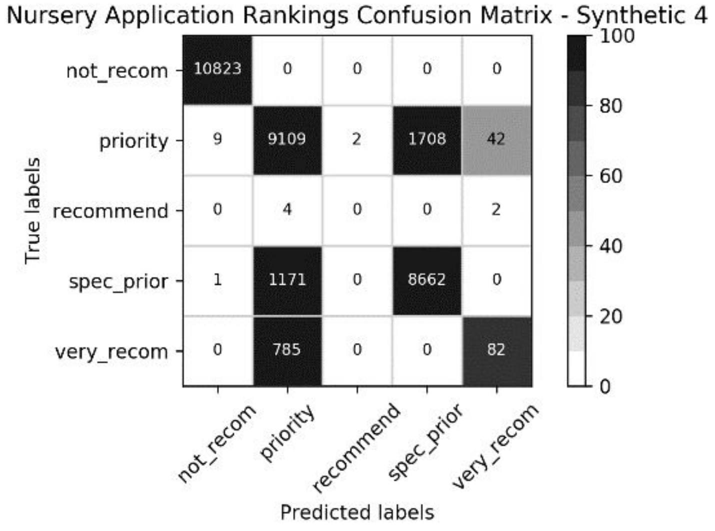
The confusion matrices for the performance of each of the five classifiers, trained on each of the eleven datasets (original, 5 synthetic non-parametric and 5 synthetic parametric) are shown in Figure 3.2.5-3.2.9 for the Linear model, Decision Tree model, KNN model, Random Forest model and SVM model respectively. The results show a higher degree of misclassification in the Nursery dataset containing categorical data, compared with the Breast Cancer dataset containing numerical data. However, this misclassification is observed in models trained with the real data and the synthetic data to a similar degree. Therefore, the issue is more likely to exist in the models used, instead of with the synthetic data used to train them. Further investigation is required to fine tune the models and try alternative, more suitable models to determine the underlying problem.



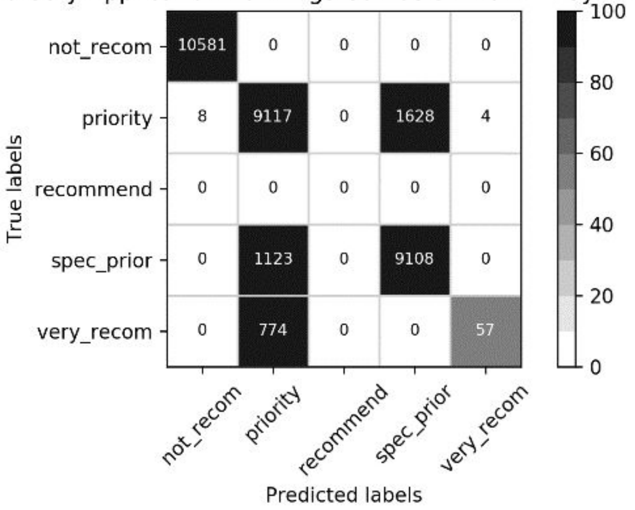



Nursery Application Rankings Confusion Matrix - Synthetic 2
True labels
not_recom
priority
recommend
spec_prior
very_recom
10883 0 0 0 0
0 9149 0 1447 18
0 7 0 0 2
0 1517 0 8534 0
0 766 0 0 77
Predicted labels
not_recom
priority
recommend
spec_prior
very_recom
100
80
60
40
20
0


Nursery Application Rankings Confusion Matrix - Synthetic 3
True labels
not_recom
priority
recommend
spec_prior
very_recom
10891 0 0 0 0
38 8802 0 1575 44
0 1 0 0 0
17 931 0 9215 0
0 799 0 0 87
Predicted labels
not_recom
priority
recommend
spec_prior
very_recom
100
80
60
40
20
0

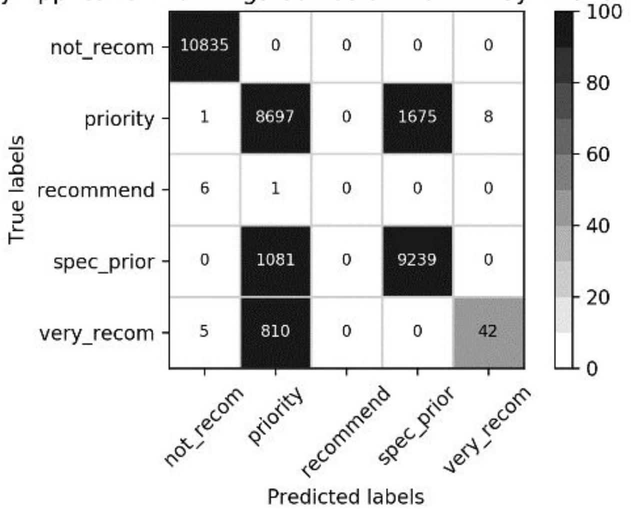

Nursery Application Rankings Confusion Matrix - Synthetic 4
True labels
not_recom
priority
recommend
spec_prior
very_recom
10823 0 0 0 0
9 9109 2 1708 42
0 4 0 0 2
1 1171 0 8662 0
0 785 0 0 82
Predicted labels
not_recom
priority
recommend
spec_prior
very_recom
100
80
60
40
20
0

Nursery Application Rankings Confusion Matrix - Synthetic 5



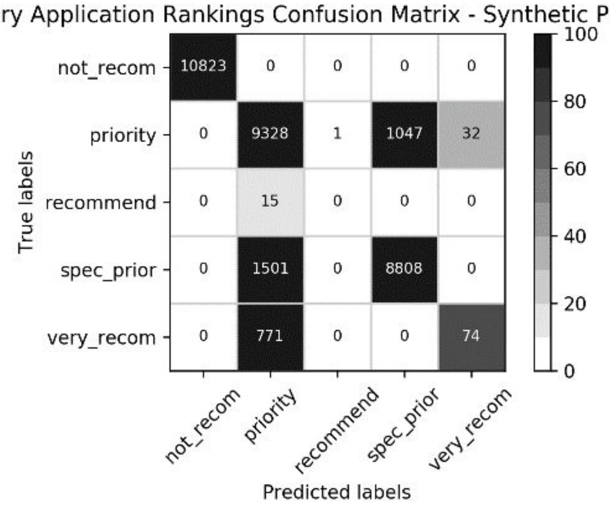
	not_recom	priority	recommend	spec_prior	very_recom
not_recom	10581	0	0	0	0
priority	8	9117	0	1628	4
recommend	0	0	0	0	0
spec_prior	0	1123	0	9108	0
very_recom	0	774	0	0	57

Nursery Application Rankings Confusion Matrix - Synthetic Parametric



	not_recom	priority	recommend	spec_prior	very_recom
not_recom	10835	0	0	0	0
priority	1	8697	0	1675	8
recommend	6	1	0	0	0
spec_prior	0	1081	0	9239	0
very_recom	5	810	0	0	42

Nursery Application Rankings Confusion Matrix - Synthetic Parametric



	not_recom	priority	recommend	spec_prior	very_recom
not_recom	10823	0	0	0	0
priority	0	9328	1	1047	32
recommend	0	15	0	0	0
spec_prior	0	1501	0	8808	0
very_recom	0	771	0	0	74

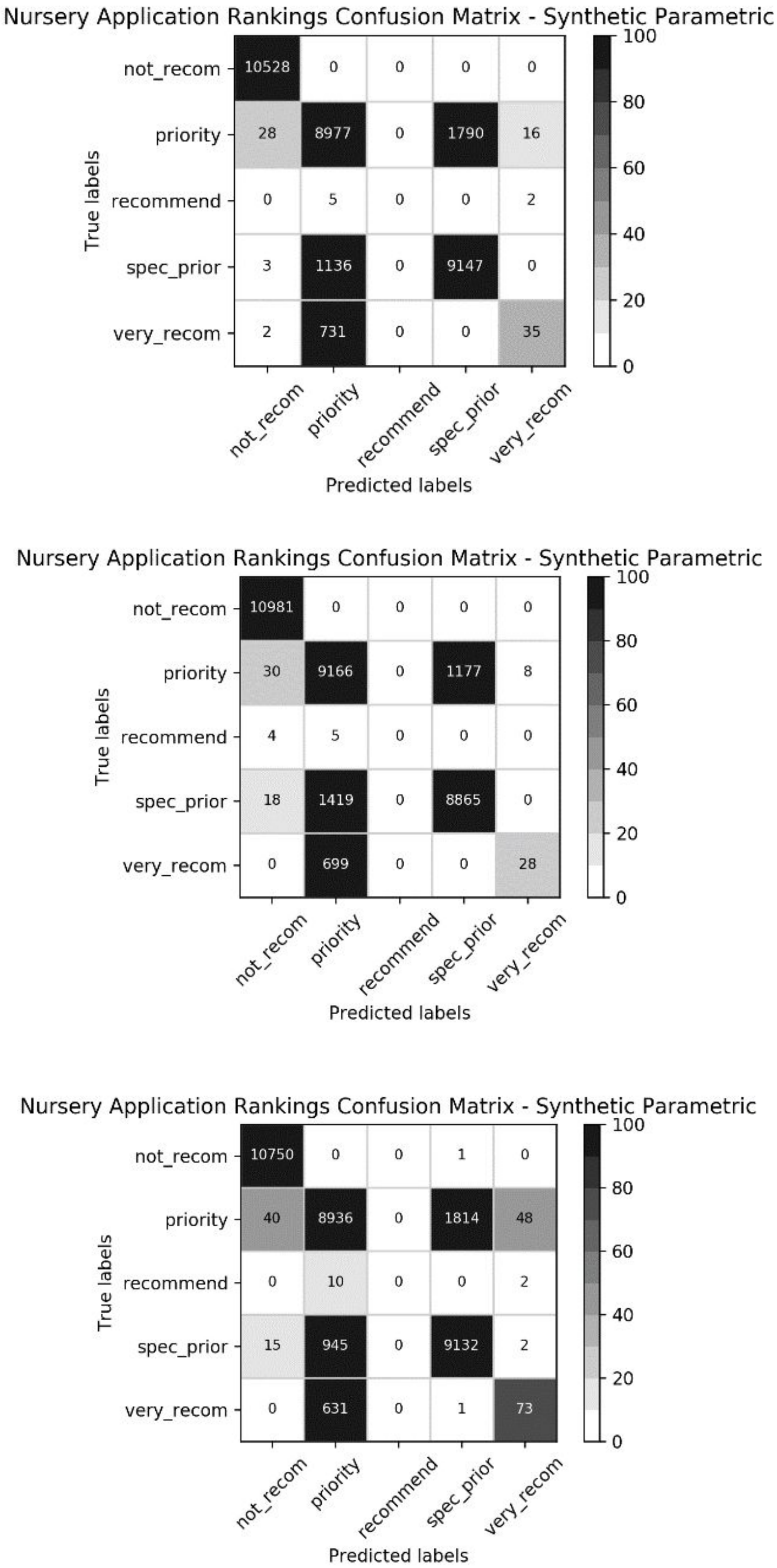
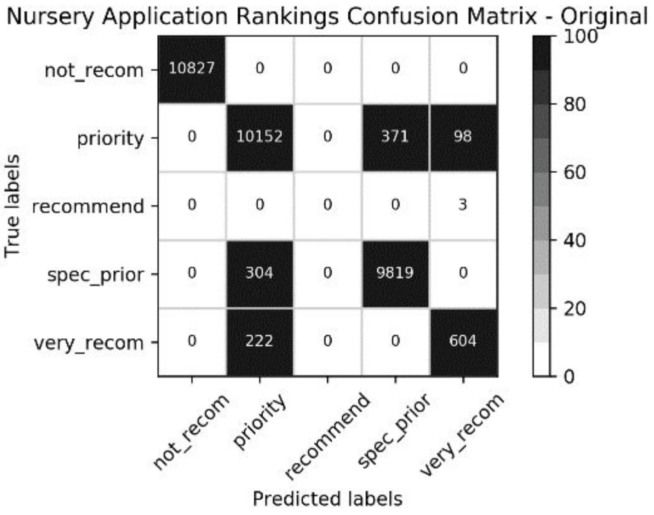
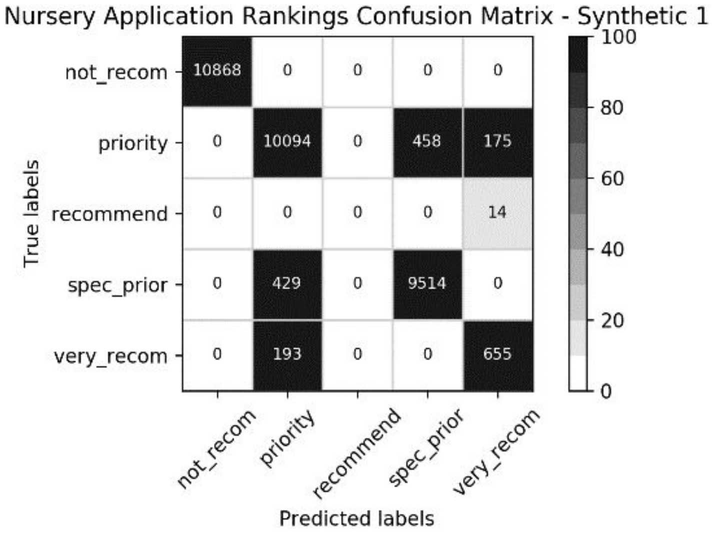
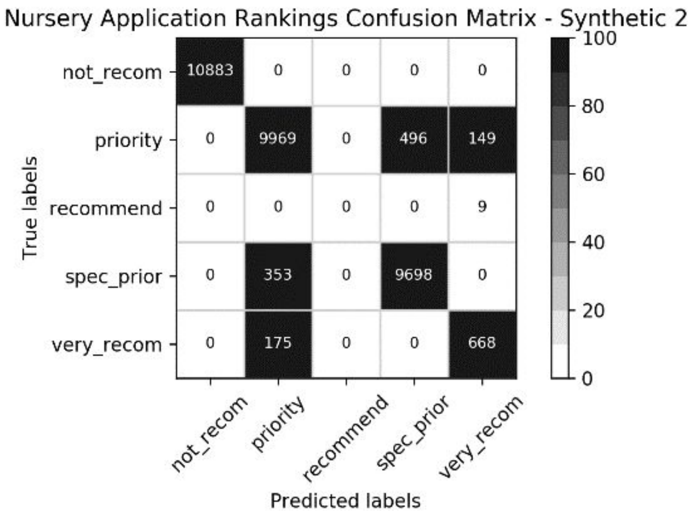
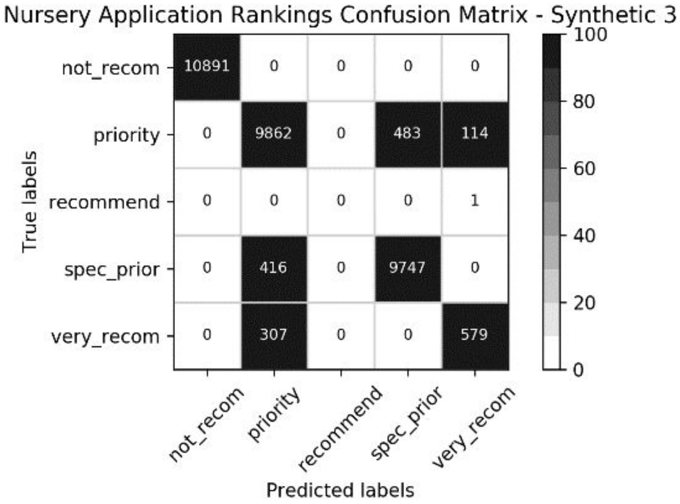


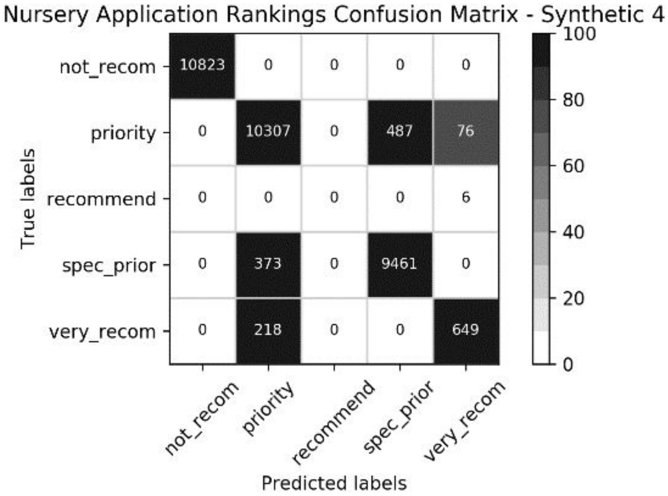
Figure 3.2.5 Confusion Matrices for the Linear Model when applied to each of the 11 datasets (1 original and 10 synthetic)

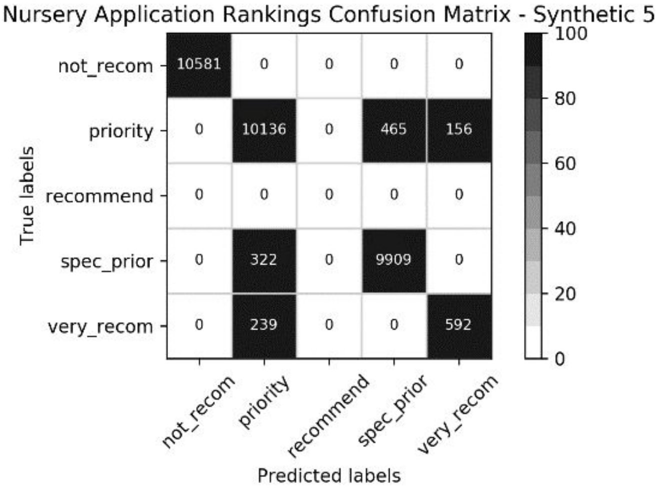

Nursery Application Rankings Confusion Matrix - Original
True labels
Predicted labels
not_recom
priority
recommend
spec_prior
very_recom
10827
0
0
0
0
0
10152
0
371
98
0
0
0
0
3
0
304
0
9819
0
0
222
0
0
604
100
80
60
40
20
0

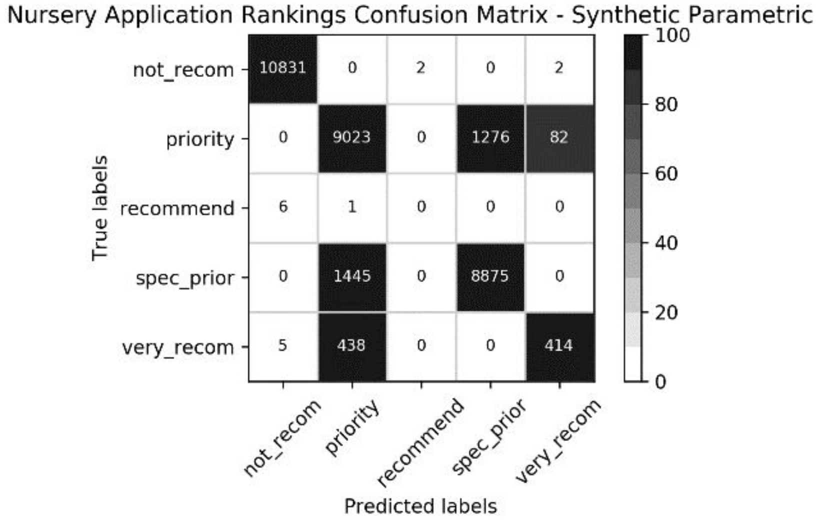

Nursery Application Rankings Confusion Matrix - Synthetic 1
True labels
Predicted labels
not_recom
priority
recommend
spec_prior
very_recom
10868
0
0
0
0
0
10094
0
458
175
0
0
0
0
14
0
429
0
9514
0
0
193
0
0
655
100
80
60
40
20
0

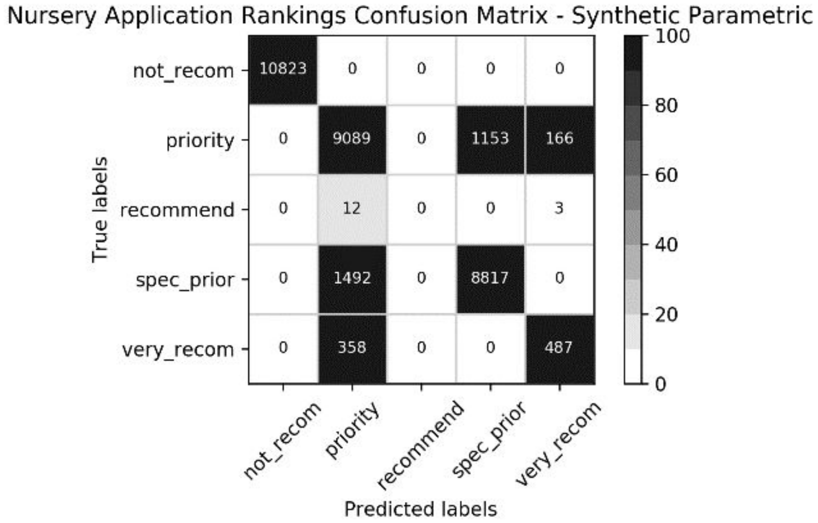

Nursery Application Rankings Confusion Matrix - Synthetic 2
True labels
Predicted labels
not_recom
priority
recommend
spec_prior
very_recom
10883
0
0
0
0
0
9969
0
496
149
0
0
0
0
9
0
353
0
9698
0
0
175
0
0
668
100
80
60
40
20
0

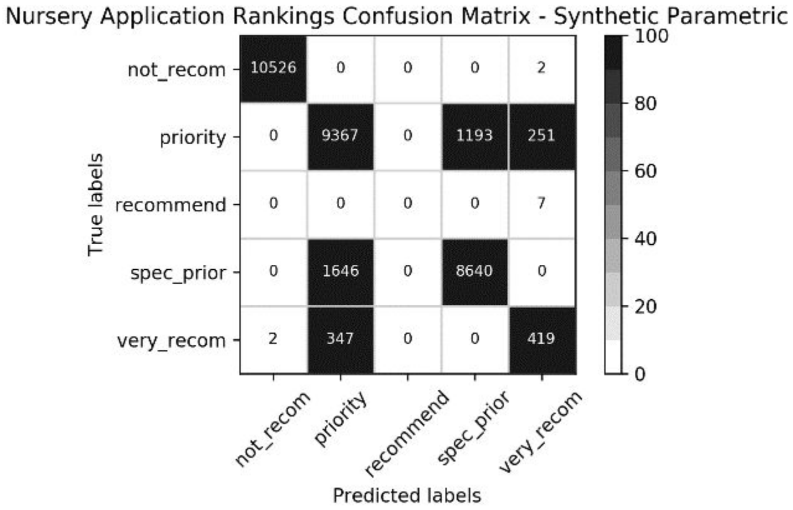

Nursery Application Rankings Confusion Matrix - Synthetic 3
True labels
Predicted labels
not_recom
priority
recommend
spec_prior
very_recom
10891 0 0 0 0
0 9862 0 483 114
0 0 0 0 1
0 416 0 9747 0
0 307 0 0 579
100
80
60
40
20
0


Nursery Application Rankings Confusion Matrix - Synthetic 4
True labels
Predicted labels
not_recom
priority
recommend
spec_prior
very_recom
10823 0 0 0 0
0 10307 0 487 76
0 0 0 0 6
0 373 0 9461 0
0 218 0 0 649
100
80
60
40
20
0


Nursery Application Rankings Confusion Matrix - Synthetic 5
True labels
Predicted labels
not_recom
priority
recommend
spec_prior
very_recom
10581 0 0 0 0
0 10136 0 465 156
0 0 0 0 0
0 322 0 9909 0
0 239 0 0 592
100
80
60
40
20
0


Nursery Application Rankings Confusion Matrix - Synthetic Parametric
True labels
not_recom 10831 0 2 0 2
priority 0 9023 0 1276 82
recommend 6 1 0 0 0
spec_prior 0 1445 0 8875 0
very_recom 5 438 0 0 414
not_recom
priority
recommend
spec_prior
very_recom
Predicted labels
100
80
60
40
20
0


Nursery Application Rankings Confusion Matrix - Synthetic Parametric
True labels
not_recom 10823 0 0 0 0
priority 0 9089 0 1153 166
recommend 0 12 0 0 3
spec_prior 0 1492 0 8817 0
very_recom 0 358 0 0 487
not_recom
priority
recommend
spec_prior
very_recom
Predicted labels
100
80
60
40
20
0


Nursery Application Rankings Confusion Matrix - Synthetic Parametric
True labels
not_recom 10526 0 0 0 2
priority 0 9367 0 1193 251
recommend 0 0 0 0 7
spec_prior 0 1646 0 8640 0
very_recom 2 347 0 0 419
not_recom
priority
recommend
spec_prior
very_recom
Predicted labels
100
80
60
40
20
0

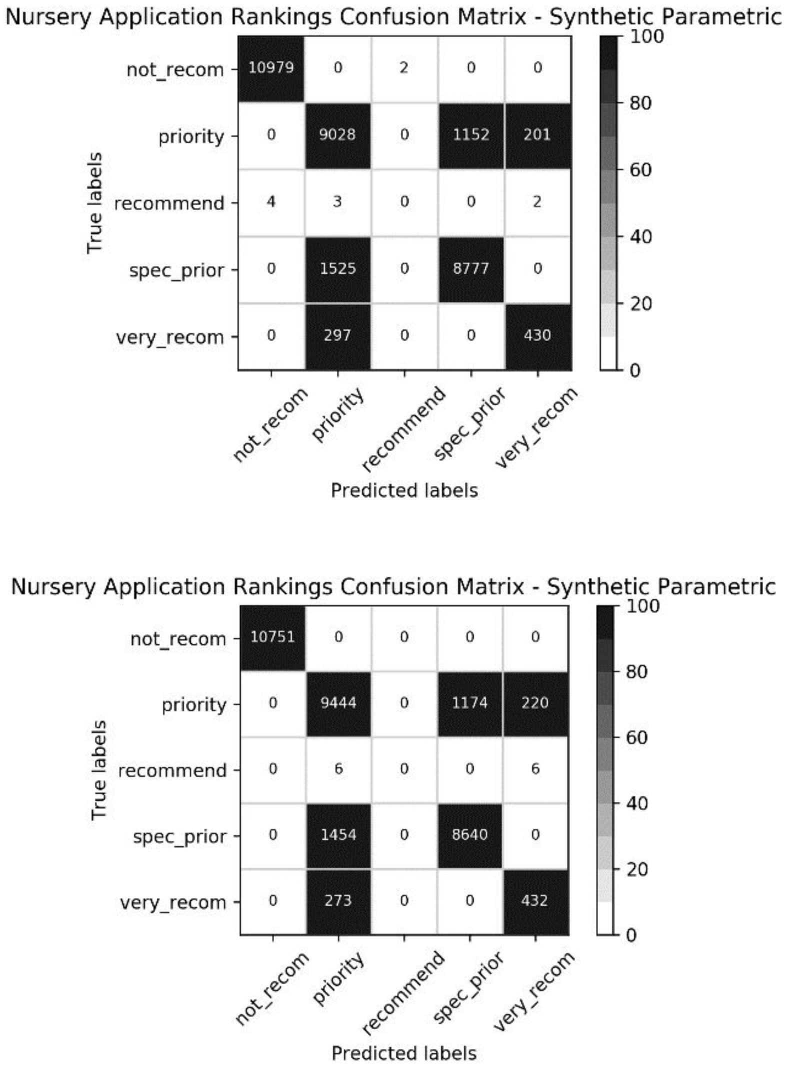
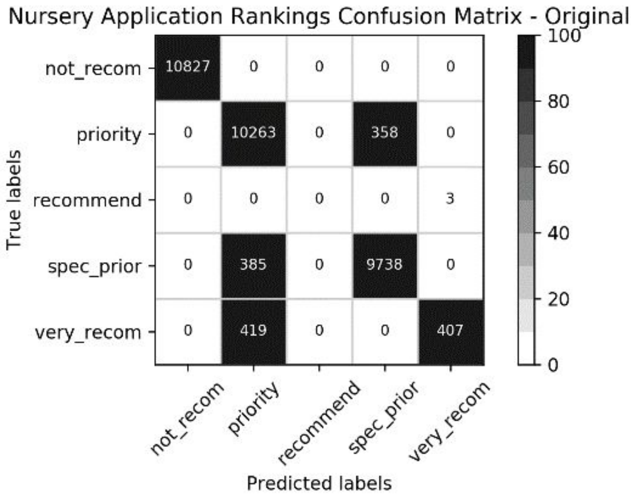
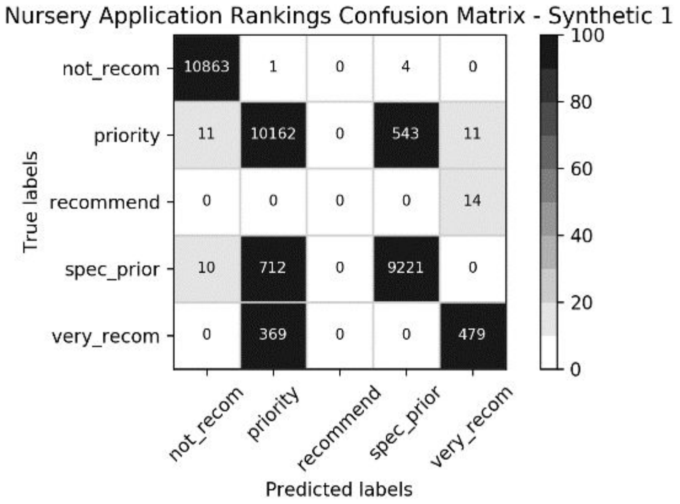
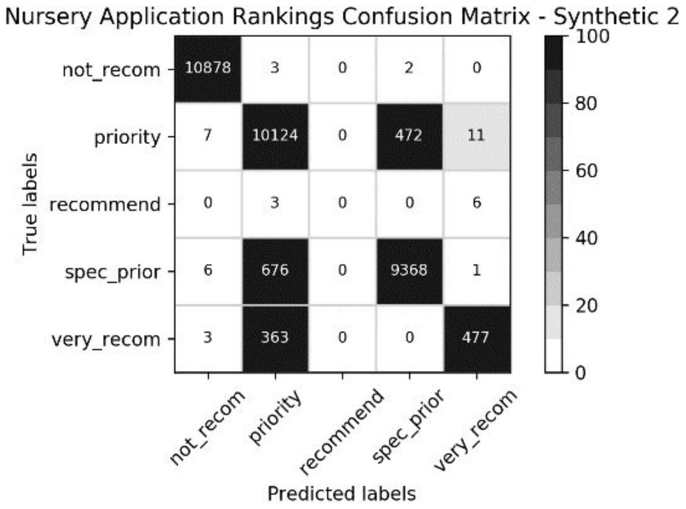
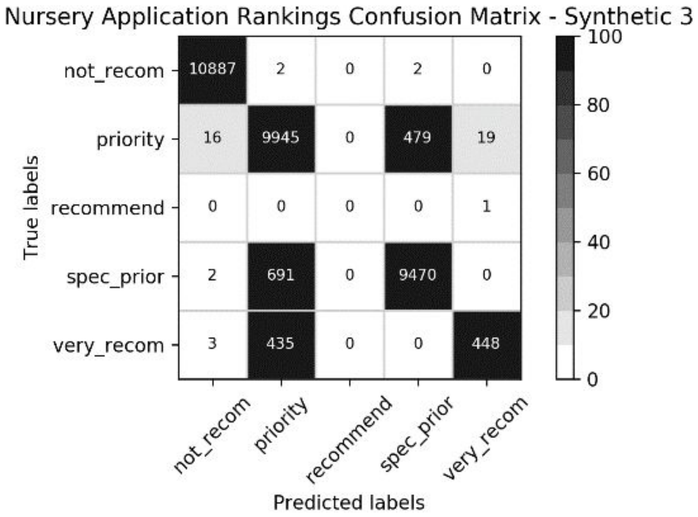


Figure 3.2.6 Confusion Matrices for the Decision Tree Model when applied to each of the 11 datasets (1 original and 10 synthetic)

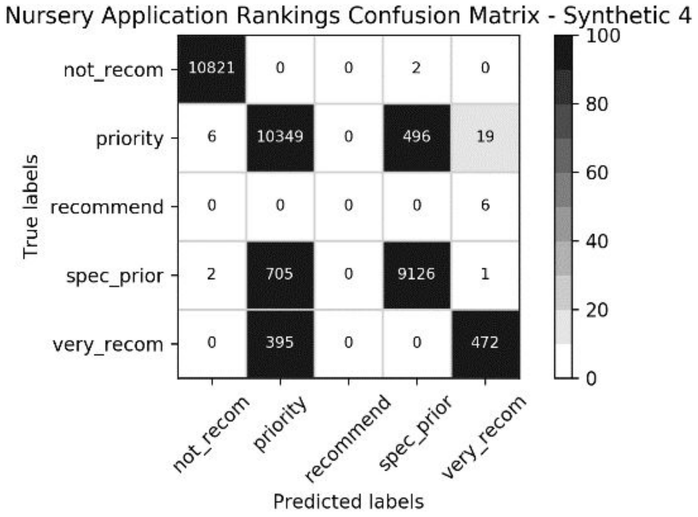

Nursery Application Rankings Confusion Matrix - Original
True labels
Predicted labels
not_recom
priority
recommend
spec_prior
very_recom
not_recom: 10827 0 0 0 0
priority: 0 10263 0 358 0
recommend: 0 0 0 0 3
spec_prior: 0 385 0 9738 0
very_recom: 0 419 0 0 407
100
80
60
40
20
0


Nursery Application Rankings Confusion Matrix - Synthetic 1
True labels
Predicted labels
not_recom
priority
recommend
spec_prior
very_recom
not_recom: 10863 1 0 4 0
priority: 11 10162 0 543 11
recommend: 0 0 0 0 14
spec_prior: 10 712 0 9221 0
very_recom: 0 369 0 0 479
100
80
60
40
20
0

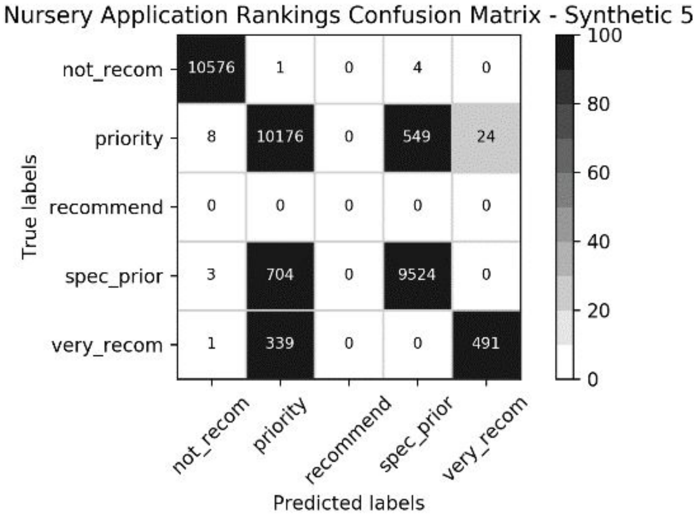

Nursery Application Rankings Confusion Matrix - Synthetic 2
True labels
Predicted labels
not_recom
priority
recommend
spec_prior
very_recom
not_recom: 10878 3 0 2 0
priority: 7 10124 0 472 11
recommend: 0 3 0 0 6
spec_prior: 6 676 0 9368 1
very_recom: 3 363 0 0 477
100
80
60
40
20
0



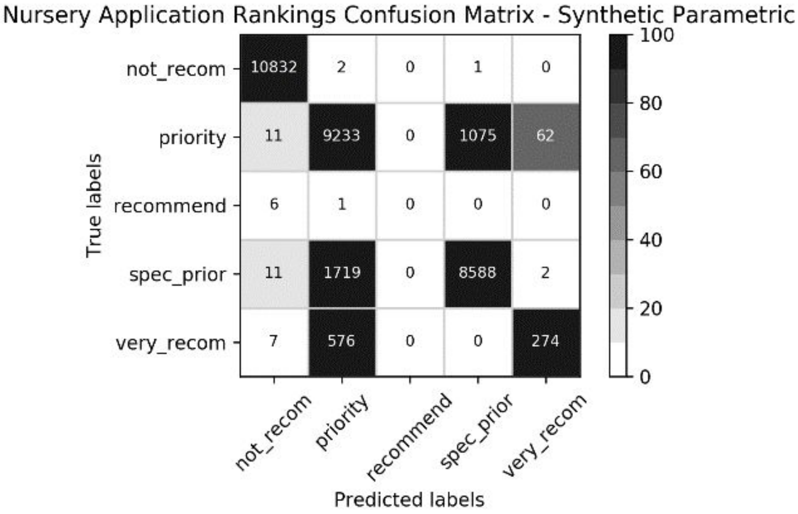
	not_recom	priority	recommend	spec_prior	very_recom
not_recom	10887	2	0	2	0
priority	16	9945	0	479	19
recommend	0	0	0	0	1
spec_prior	2	691	0	9470	0
very_recom	3	435	0	0	448

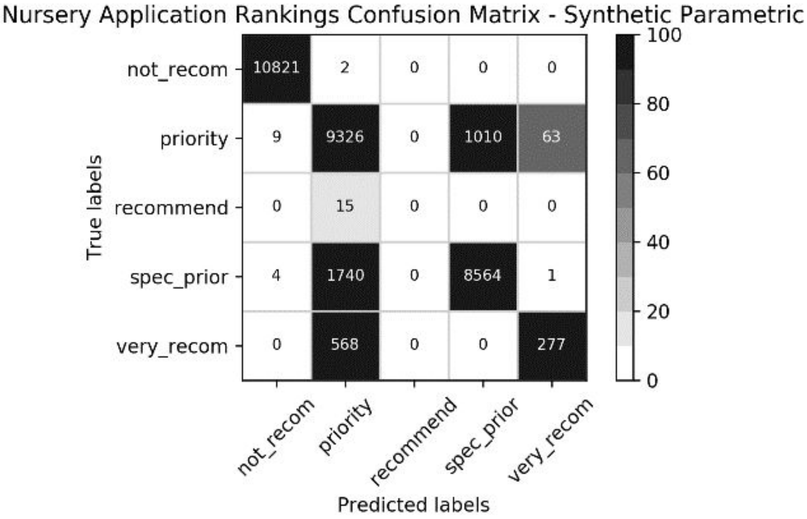


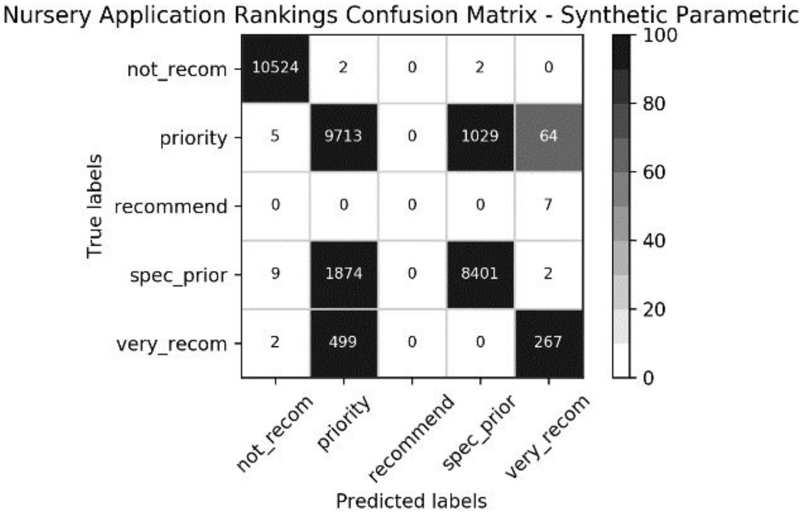
	not_recom	priority	recommend	spec_prior	very_recom
not_recom	10821	0	0	2	0
priority	6	10349	0	496	19
recommend	0	0	0	0	6
spec_prior	2	705	0	9126	1
very_recom	0	395	0	0	472

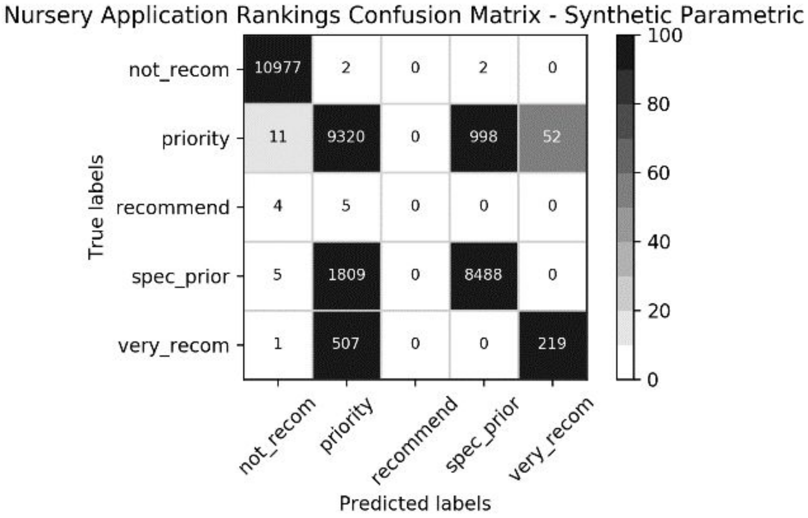


	not_recom	priority	recommend	spec_prior	very_recom
not_recom	10576	1	0	4	0
priority	8	10176	0	549	24
recommend	0	0	0	0	0
spec_prior	3	704	0	9524	0
very_recom	1	339	0	0	491


Nursery Application Rankings Confusion Matrix - Synthetic Parametric
True labels
not_recom
priority
recommend
spec_prior
very_recom
10832
2
0
1
0
11
9233
0
1075
62
6
1
0
0
0
11
1719
0
8588
2
7
576
0
0
274
Predicted labels
not_recom
priority
recommend
spec_prior
very_recom
100
80
60
40
20
0


Nursery Application Rankings Confusion Matrix - Synthetic Parametric
True labels
not_recom
priority
recommend
spec_prior
very_recom
10821
2
0
0
0
9
9326
0
1010
63
0
15
0
0
0
4
1740
0
8564
1
0
568
0
0
277
Predicted labels
not_recom
priority
recommend
spec_prior
very_recom
100
80
60
40
20
0


Nursery Application Rankings Confusion Matrix - Synthetic Parametric
True labels
not_recom
priority
recommend
spec_prior
very_recom
10524
2
0
2
0
5
9713
0
1029
64
0
0
0
0
7
9
1874
0
8401
2
2
499
0
0
267
Predicted labels
not_recom
priority
recommend
spec_prior
very_recom
100
80
60
40
20
0



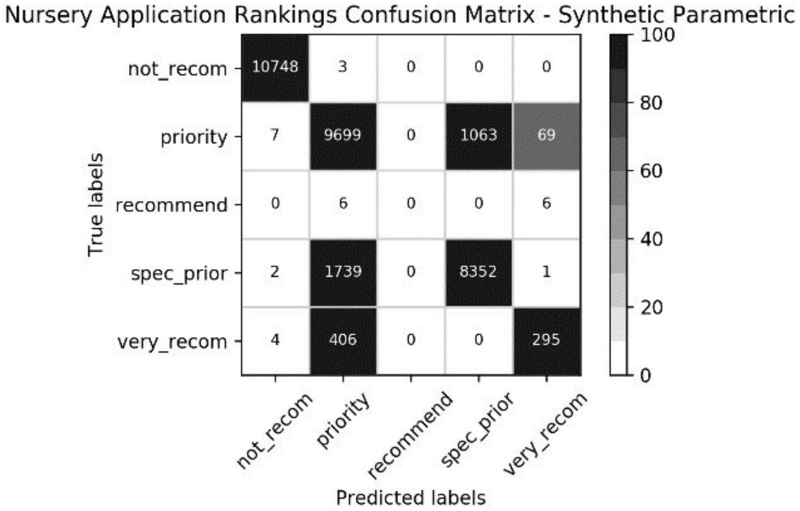
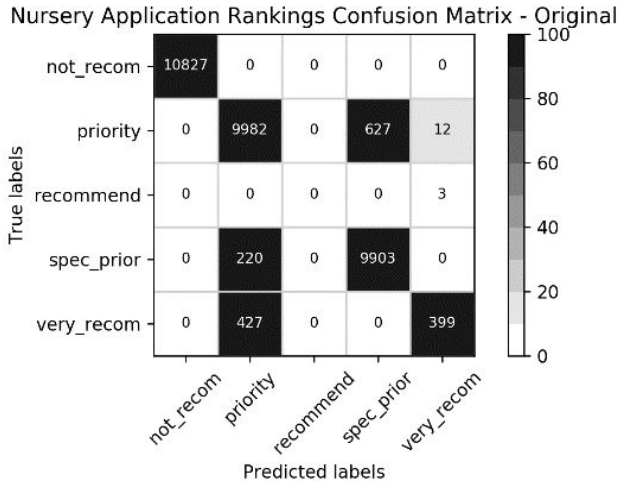
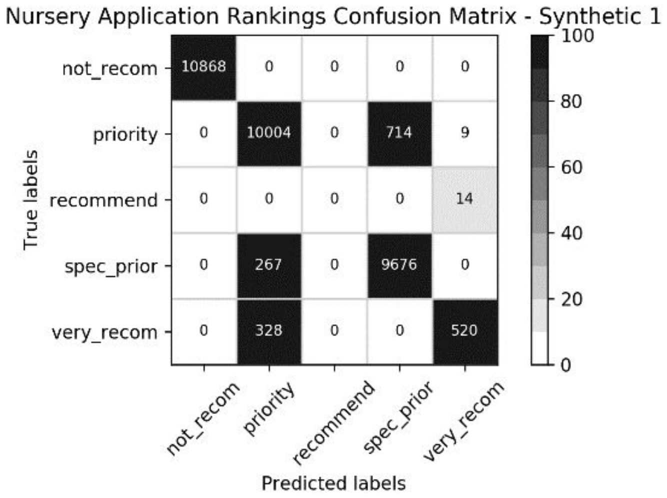
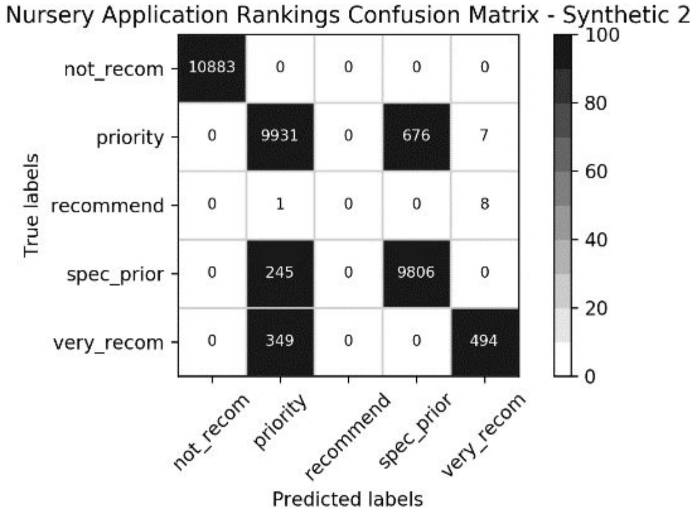
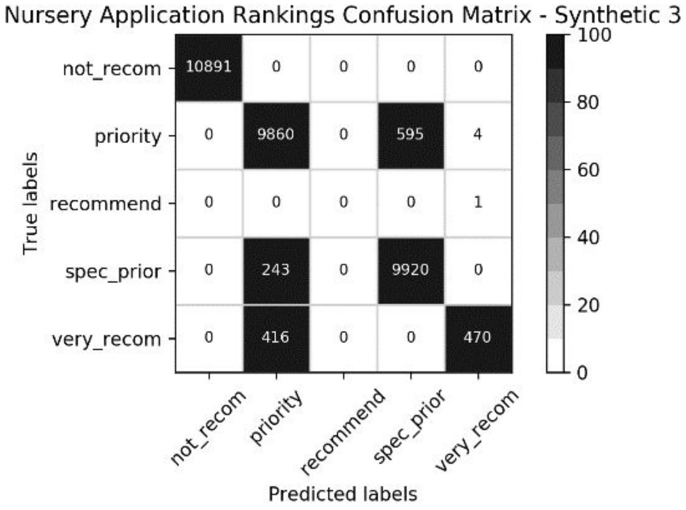


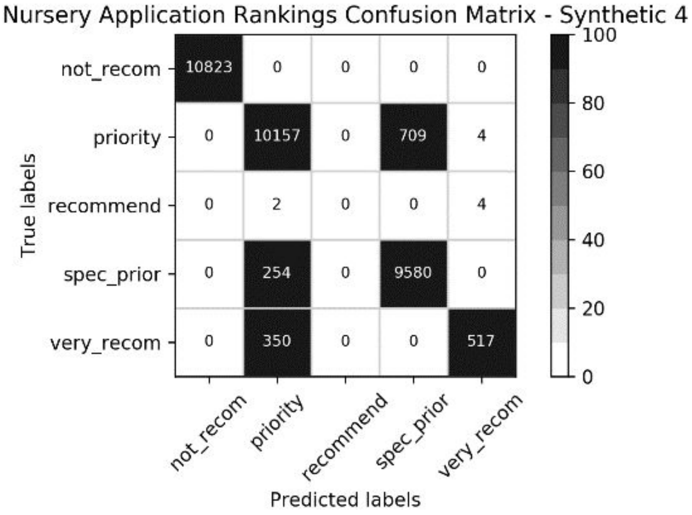
Figure 3.2.7 Confusion Matrices for the KNN Model when applied to each of the 11 datasets (1 original and 10 synthetic)

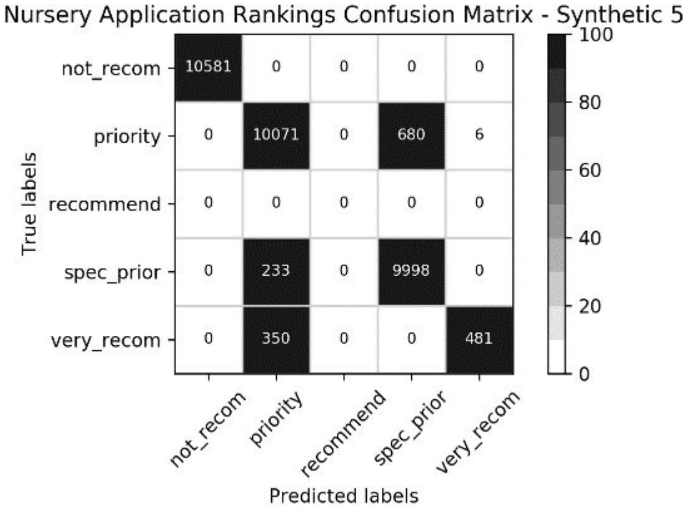

Nursery Application Rankings Confusion Matrix - Original
True labels
Predicted labels
not_recom
priority
recommend
spec_prior
very_recom
10827 0 0 0 0
0 9982 0 627 12
0 0 0 0 3
0 220 0 9903 0
0 427 0 0 399
100
80
60
40
20
0

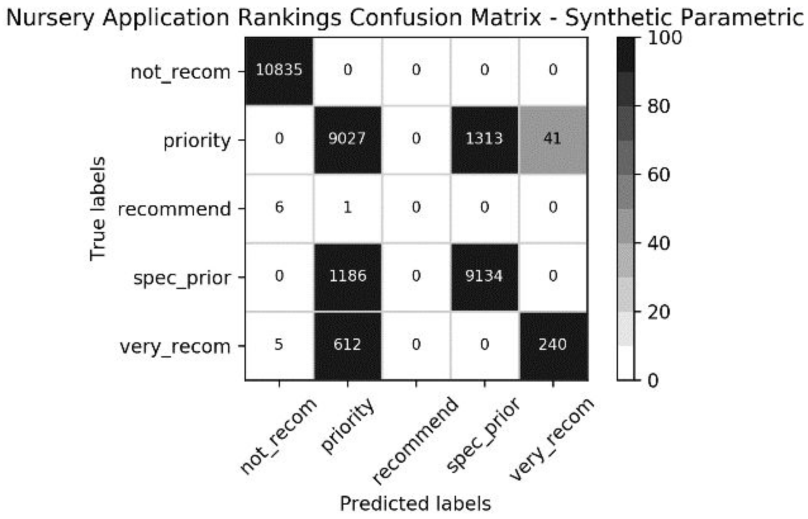

Nursery Application Rankings Confusion Matrix - Synthetic 1
True labels
Predicted labels
not_recom
priority
recommend
spec_prior
very_recom
10868 0 0 0 0
0 10004 0 714 9
0 0 0 0 14
0 267 0 9676 0
0 328 0 0 520
100
80
60
40
20
0

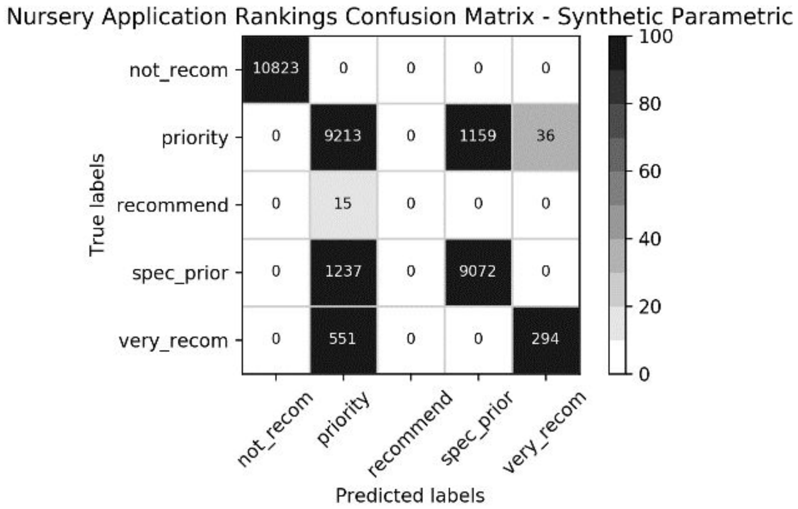

Nursery Application Rankings Confusion Matrix - Synthetic 2
True labels
Predicted labels
not_recom
priority
recommend
spec_prior
very_recom
10883 0 0 0 0
0 9931 0 676 7
0 1 0 0 8
0 245 0 9806 0
0 349 0 0 494
100
80
60
40
20
0

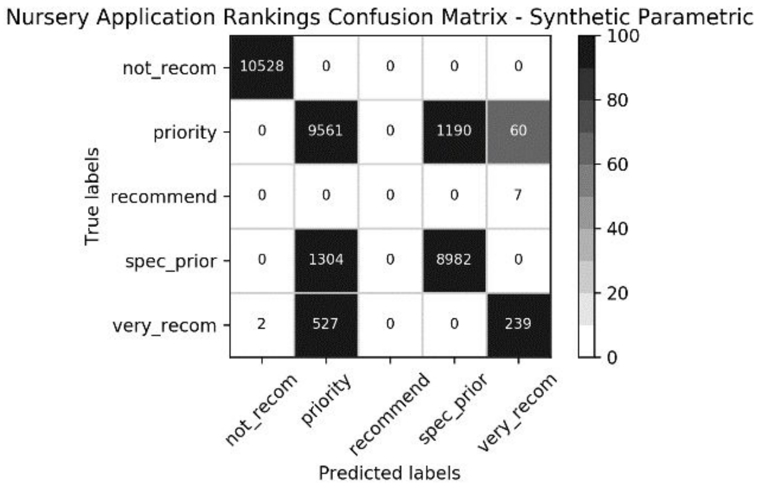

Nursery Application Rankings Confusion Matrix - Synthetic 3
True labels
not_recom
priority
recommend
spec_prior
very_recom
10891
0
0
0
0
0
9860
0
595
4
0
0
0
0
1
0
243
0
9920
0
0
416
0
0
470
not_recom
priority
recommend
spec_prior
very_recom
Predicted labels
100
80
60
40
20
0


Nursery Application Rankings Confusion Matrix - Synthetic 4
True labels
not_recom
priority
recommend
spec_prior
very_recom
10823
0
0
0
0
0
10157
0
709
4
0
2
0
0
4
0
254
0
9580
0
0
350
0
0
517
not_recom
priority
recommend
spec_prior
very_recom
Predicted labels
100
80
60
40
20
0


Nursery Application Rankings Confusion Matrix - Synthetic 5
True labels
not_recom
priority
recommend
spec_prior
very_recom
10581
0
0
0
0
0
10071
0
680
6
0
0
0
0
0
0
233
0
9998
0
0
350
0
0
481
not_recom
priority
recommend
spec_prior
very_recom
Predicted labels
100
80
60
40
20
0


Nursery Application Rankings Confusion Matrix - Synthetic Parametric
True labels
not_recom
priority
recommend
spec_prior
very_recom
10835 0 0 0 0
0 9027 0 1313 41
6 1 0 0 0
0 1186 0 9134 0
5 612 0 0 240
Predicted labels
not_recom
priority
recommend
spec_prior
very_recom
100
80
60
40
20
0


Nursery Application Rankings Confusion Matrix - Synthetic Parametric
True labels
not_recom
priority
recommend
spec_prior
very_recom
10823 0 0 0 0
0 9213 0 1159 36
0 15 0 0 0
0 1237 0 9072 0
0 551 0 0 294
Predicted labels
not_recom
priority
recommend
spec_prior
very_recom
100
80
60
40
20
0


Nursery Application Rankings Confusion Matrix - Synthetic Parametric
True labels
not_recom
priority
recommend
spec_prior
very_recom
10528 0 0 0 0
0 9561 0 1190 60
0 0 0 0 7
0 1304 0 8982 0
2 527 0 0 239
Predicted labels
not_recom
priority
recommend
spec_prior
very_recom
100
80
60
40
20
0

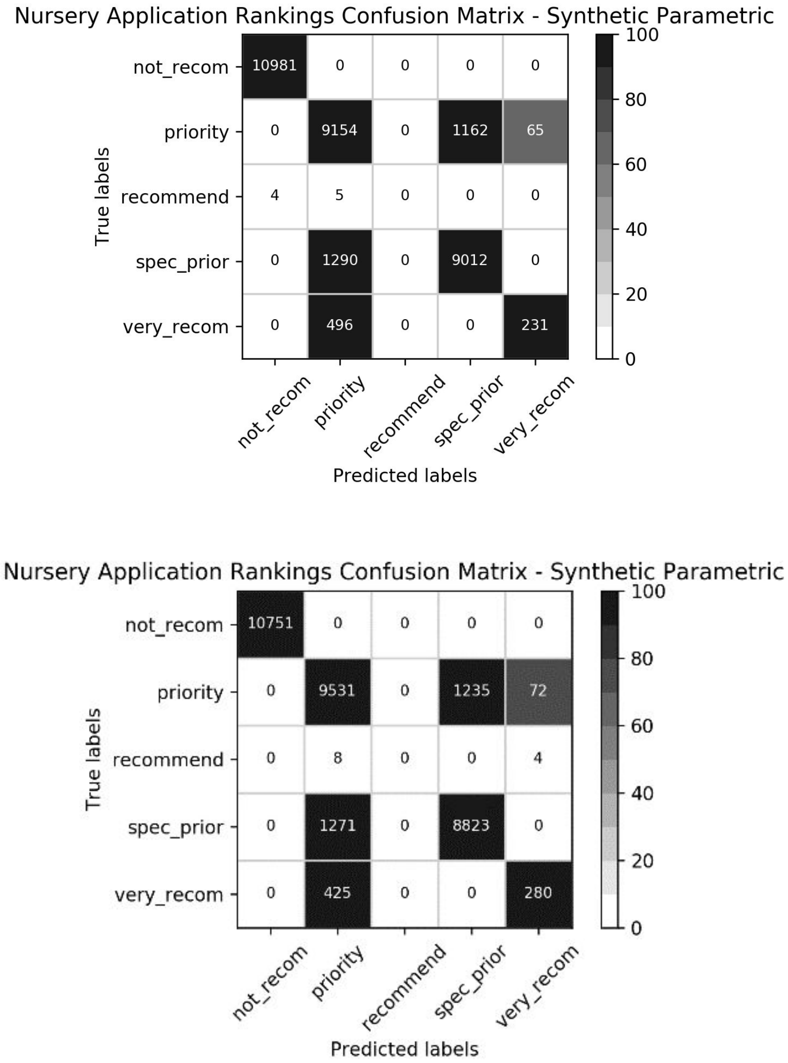
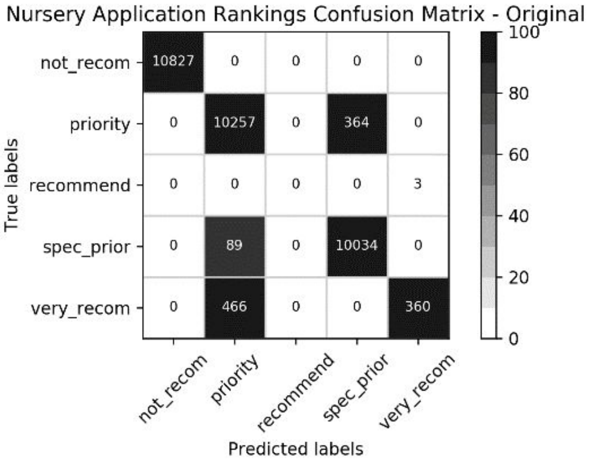
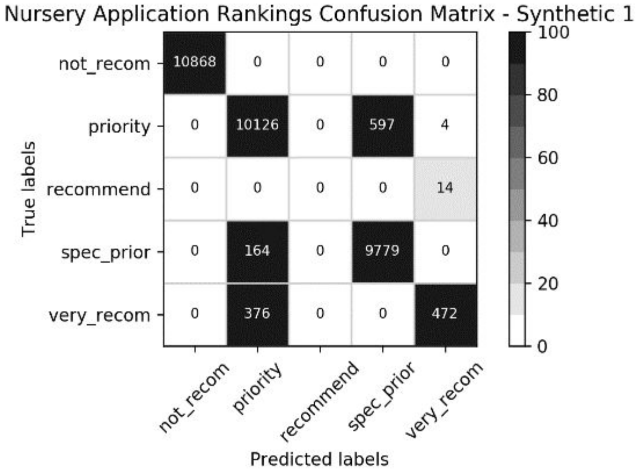


Figure 3.2.8 Confusion Matrices for the Random Forest Model when applied to each of the 11 datasets (1 original and 10 synthetic)



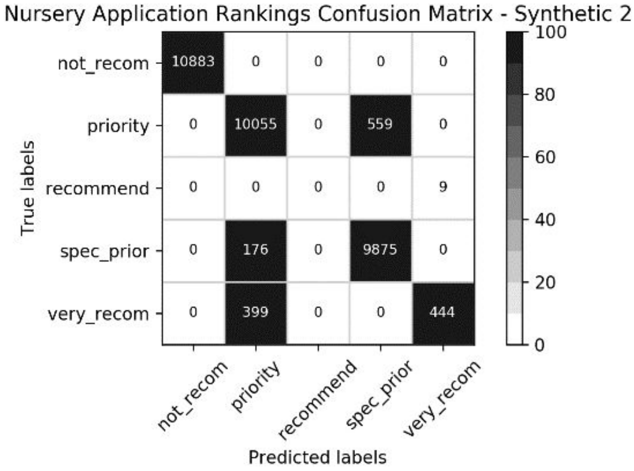
Nursery Application Rankings Confusion Matrix - Original

True labels \ Predicted labels	not_recom	priority	recommend	spec_prior	very_recom
not_recom	10827	0	0	0	0
priority	0	10257	0	364	0
recommend	0	0	0	0	3
spec_prior	0	89	0	10034	0
very_recom	0	466	0	0	360



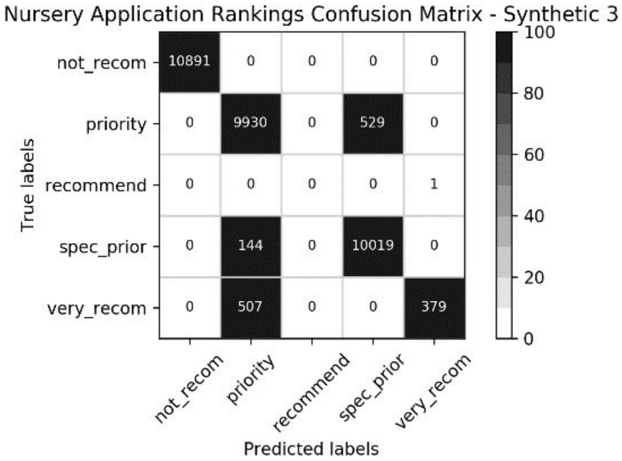
Nursery Application Rankings Confusion Matrix - Synthetic 1

True labels \ Predicted labels	not_recom	priority	recommend	spec_prior	very_recom
not_recom	10868	0	0	0	0
priority	0	10126	0	597	4
recommend	0	0	0	0	14
spec_prior	0	164	0	9779	0
very_recom	0	376	0	0	472

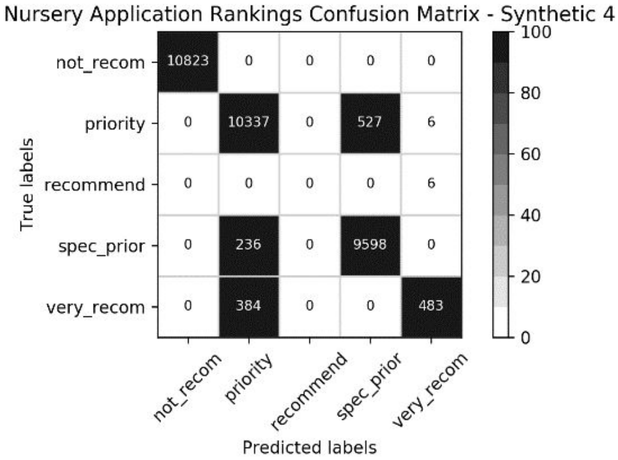


Nursery Application Rankings Confusion Matrix - Synthetic 2

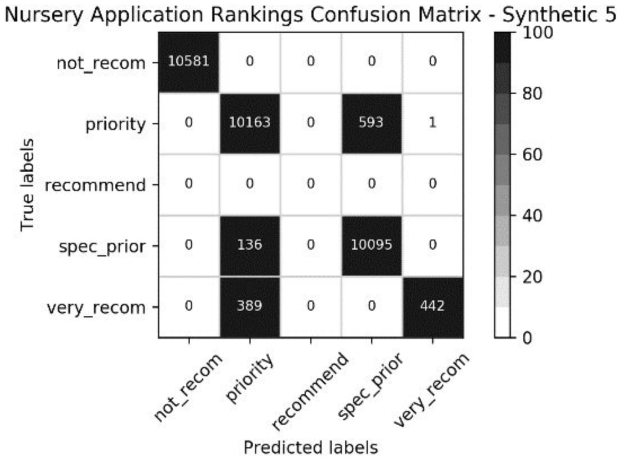
True labels \ Predicted labels	not_recom	priority	recommend	spec_prior	very_recom
not_recom	10883	0	0	0	0
priority	0	10055	0	559	0
recommend	0	0	0	0	9
spec_prior	0	176	0	9875	0
very_recom	0	399	0	0	444



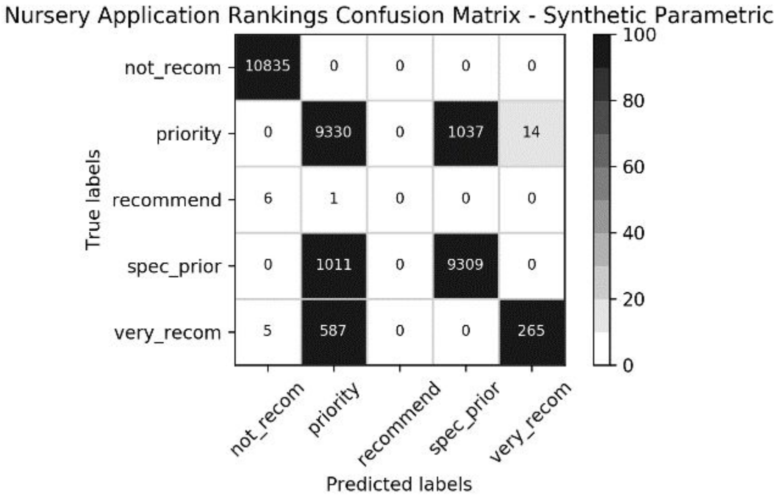
	not_recom	priority	recommend	spec_prior	very_recom
not_recom	10891	0	0	0	0
priority	0	9930	0	529	0
recommend	0	0	0	0	1
spec_prior	0	144	0	10019	0
very_recom	0	507	0	0	379

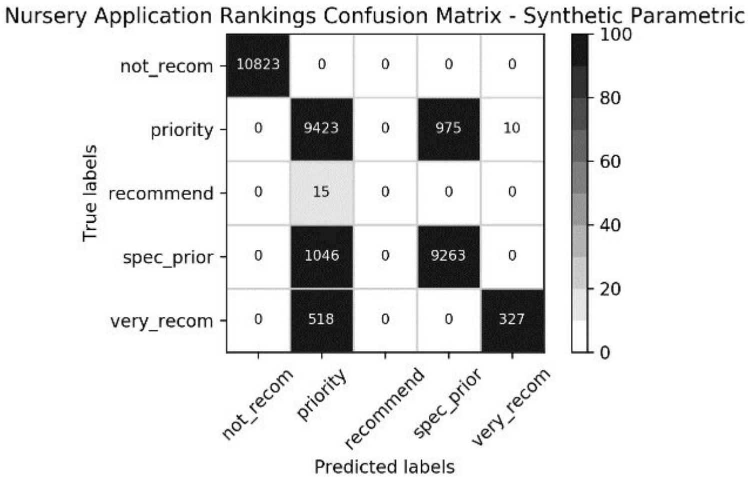


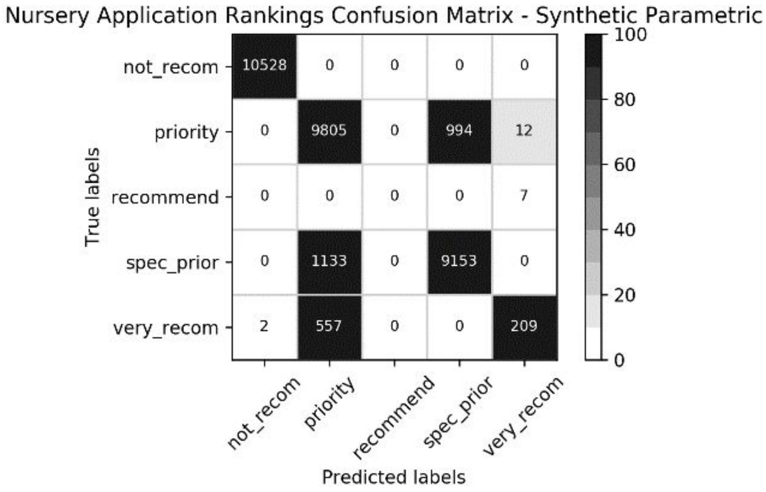
	not_recom	priority	recommend	spec_prior	very_recom
not_recom	10823	0	0	0	0
priority	0	10337	0	527	6
recommend	0	0	0	0	6
spec_prior	0	236	0	9598	0
very_recom	0	384	0	0	483

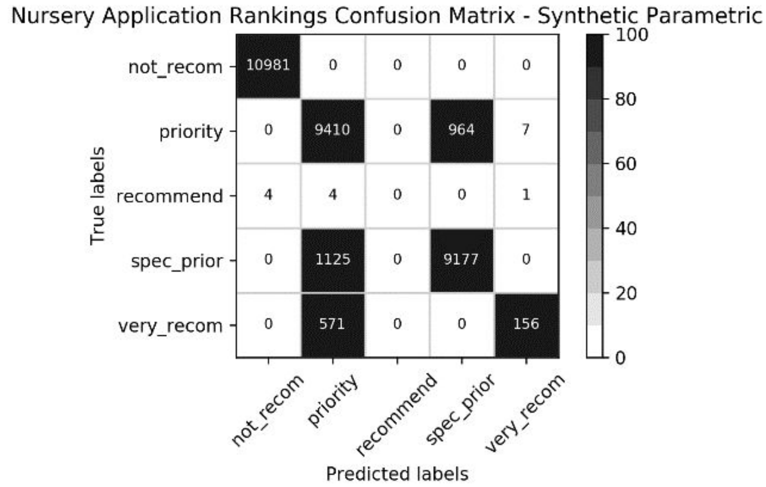


	not_recom	priority	recommend	spec_prior	very_recom
not_recom	10581	0	0	0	0
priority	0	10163	0	593	1
recommend	0	0	0	0	0
spec_prior	0	136	0	10095	0
very_recom	0	389	0	0	442


Nursery Application Rankings Confusion Matrix - Synthetic Parametric
True labels
not_recom
priority
recommend
spec_prior
very_recom
10835 0 0 0 0
0 9330 0 1037 14
6 1 0 0 0
0 1011 0 9309 0
5 587 0 0 265
not_recom
priority
recommend
spec_prior
very_recom
Predicted labels
100
80
60
40
20
0


Nursery Application Rankings Confusion Matrix - Synthetic Parametric
True labels
not_recom
priority
recommend
spec_prior
very_recom
10823 0 0 0 0
0 9423 0 975 10
0 15 0 0 0
0 1046 0 9263 0
0 518 0 0 327
not_recom
priority
recommend
spec_prior
very_recom
Predicted labels
100
80
60
40
20
0


Nursery Application Rankings Confusion Matrix - Synthetic Parametric
True labels
not_recom
priority
recommend
spec_prior
very_recom
10528 0 0 0 0
0 9805 0 994 12
0 0 0 0 7
0 1133 0 9153 0
2 557 0 0 209
not_recom
priority
recommend
spec_prior
very_recom
Predicted labels
100
80
60
40
20
0



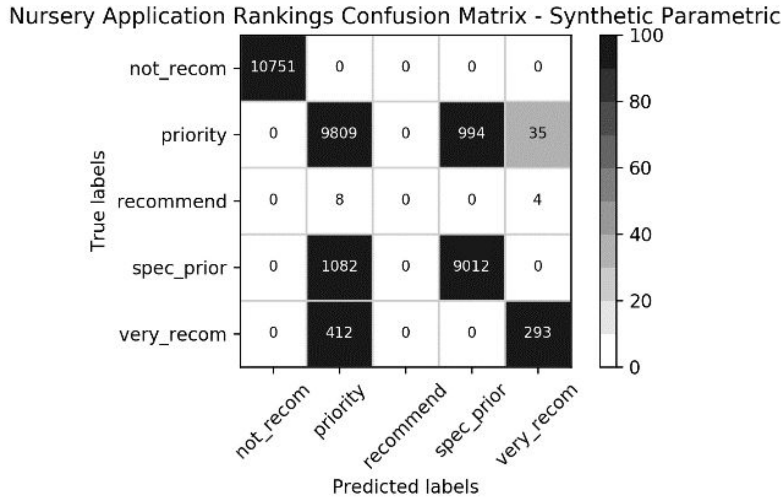


Figure 3.2.9 Confusion Matrices for the SVM Model when applied to each of the 11 datasets (1 original and 10 synthetic)

Nursery Dataset Cross Comparison

A cross comparison was also carried out for the datasets synthesised from the Nursery dataset as well as the original data to determine how well classifiers trained on synthetic data would perform when presented with real data. In this example the training dataset comprises 100% of a synthetic dataset and the test set for each comprises 100% of the original dataset. The training dataset comprises 100% of the dataset listed in column 1 of Table 3.2.5 and the test set for each comprises 100% of the original dataset. Table 3.2.5 illustrates the accuracy scores. We observe high accuracy across all models trained on all synthetic datasets and tested on the real data, however in this case, non-parametric synthetic data outperforms parametric synthetic data, and the SVM and linear models produce the lowest accuracy, whilst SVM achieves the highest average accuracy as per all previous results.

Table 3.2.5 Comparison of accuracy scores achieved by each model when trained with 100% of the dataset listed in column one and tested with 100% of the original dataset.

Dataset	Linear Model	Decision Tree	KNN	Random Forest	SVM
Original	0.911	0.976	0.949	0.975	**0.980**
Synthetic Non-Parametric V1	0.898	0.966	0.960	0.961	**0.979**
Synthetic Non-Parametric V2	0.903	0.964	0.958	0.965	**0.979**
Synthetic Non-Parametric V3	0.910	0.965	0.957	0.965	**0.973**
Synthetic Non-Parametric V4	0.887	0.966	0.957	0.964	**0.976**
Synthetic Non-Parametric V5	0.912	0.965	0.954	0.961	**0.976**
Synthetic Parametric V1	0.907	0.915	0.899	0.911	**0.921**
Synthetic Parametric V2	0.902	0.909	0.902	0.920	**0.927**
Synthetic Parametric V3	0.887	0.909	0.902	0.917	**0.924**
Synthetic Parametric V4	0.897	0.909	0.896	0.915	**0.925**
Synthetic Parametric V5	0.895	0.905	0.899	0.914	**0.923**

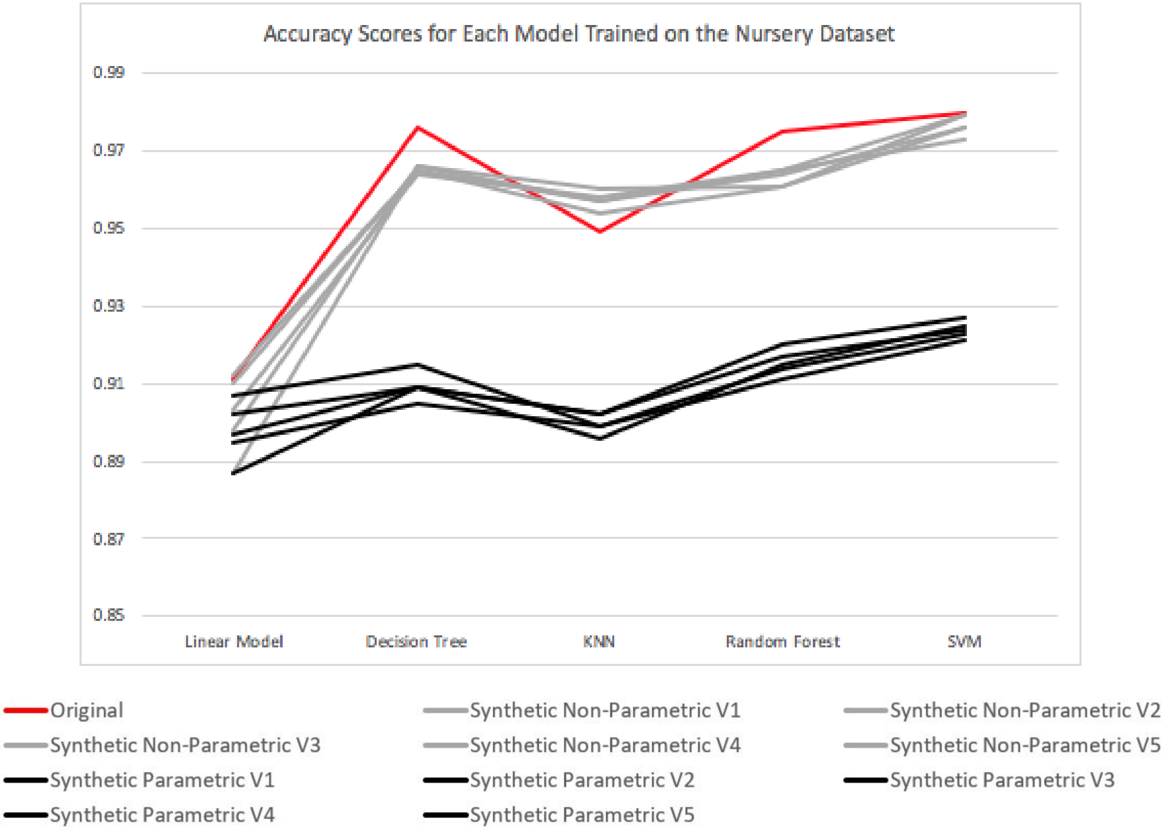


Figure 3.2.10 Comparison of accuracy scores achieved by each model when trained with 100% of the dataset listed in column one and tested with 100% of the original dataset.

Tables 3.2.6-3.2.8 illustrate the precision, recall and F1 measures, respectively for each of the five classification models after being trained by each dataset and tested with the original dataset. In cross comparisons, precision, recall and F1 scores are lower than accuracy scores however the trend in model performance is similar.

Table 3.2.6 Comparison of precision scores achieved by each model when trained with 100% of the dataset listed in column one and tested with 100% of the original dataset.

Dataset	Linear Model	Decision Tree	KNN	Random Forest	SVM
Original	0.718	0.777	0.772	0.781	0.786
Synthetic Non-Parametric V1	0.541	0.734	0.764	0.770	0.785
Synthetic Non-Parametric V2	0.701	0.736	0.767	0.773	0.785
Synthetic Non-Parametric V3	0.643	0.753	0.763	0.776	0.781
Synthetic Non-Parametric V4	0.736	0.753	0.764	0.776	0.780
Synthetic Non-Parametric V5	0.713	0.744	0.764	0.771	0.782
Synthetic Parametric V1	0.670	0.697	0.705	0.723	0.729
Synthetic Parametric V2	0.543	0.692	0.713	0.735	0.739

Synthetic Parametric V3	0.700	0.684	0.709	0.726	0.733
Synthetic Parametric V4	0.541	0.691	0.703	0.723	0.744
Synthetic Parametric V5	0.640	0.684	0.714	0.711	0.736

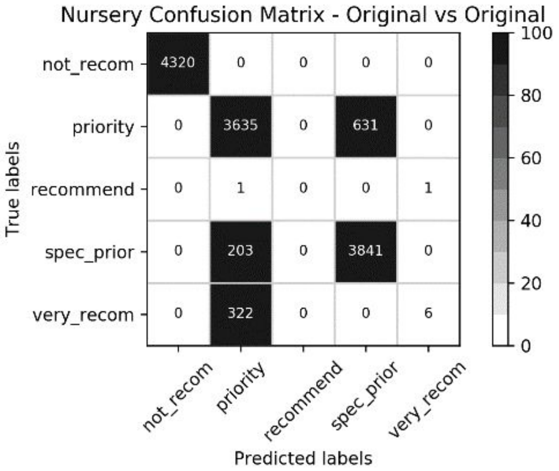
Table 3.2.7 Comparison of recall scores achieved by each model when trained with 100% of the dataset listed in column one and tested with 100% of the original dataset.

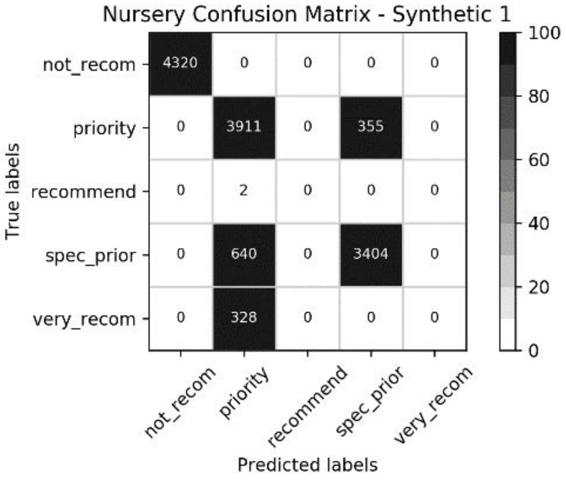
Dataset	**Linear Model**	**Decision Tree**	**KNN**	**Random Forest**	**SVM**
Original	0.564	0.721	0.748	0.719	0.718
Synthetic Non-Parametric V1	0.552	0.737	0.701	0.700	0.740
Synthetic Non-Parametric V2	0.580	0.746	0.708	0.718	0.732
Synthetic Non-Parametric V3	0.566	0.719	0.696	0.704	0.716
Synthetic Non-Parametric V4	0.546	0.726	0.698	0.699	0.724
Synthetic Non-Parametric V5	0.591	0.719	0.696	0.696	0.727
Synthetic Parametric V1	0.561	0.655	0.624	0.622	0.657
Synthetic Parametric V2	0.555	0.676	0.627	0.641	0.663
Synthetic Parametric V3	0.604	0.669	0.637	0.643	0.661
Synthetic Parametric V4	0.551	0.655	0.613	0.630	0.649
Synthetic Parametric V5	0.555	0.668	0.634	0.639	0.658

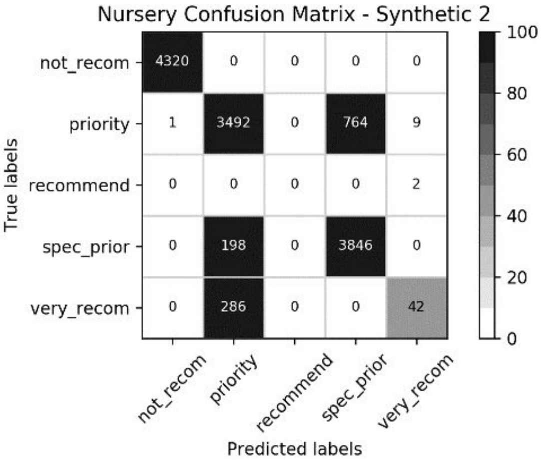
Table 3.2.8 Comparison of f1 scores achieved by each model when trained with 100% of the dataset listed in column one and tested with 100% of the original dataset.

Dataset	**Linear Model**	**Decision Tree**	**KNN**	**Random Forest**	**SVM**
Original	0.560	0.743	0.757	0.743	0.744
Synthetic Non-Parametric V1	0.546	0.736	0.725	0.725	0.759
Synthetic Non-Parametric V2	0.591	0.741	0.731	0.740	0.754
Synthetic Non-Parametric V3	0.565	0.734	0.721	0.730	0.740
Synthetic Non-Parametric V4	0.540	0.738	0.723	0.727	0.746
Synthetic Non-Parametric V5	0.606	0.730	0.721	0.723	0.749
Synthetic Parametric V1	0.557	0.671	0.648	0.648	0.682
Synthetic Parametric V2	0.548	0.684	0.653	0.668	0.689
Synthetic Parametric V3	0.625	0.676	0.661	0.669	0.686
Synthetic Parametric V4	0.545	0.670	0.638	0.655	0.678
Synthetic Parametric V5	0.554	0.675	0.659	0.663	0.684

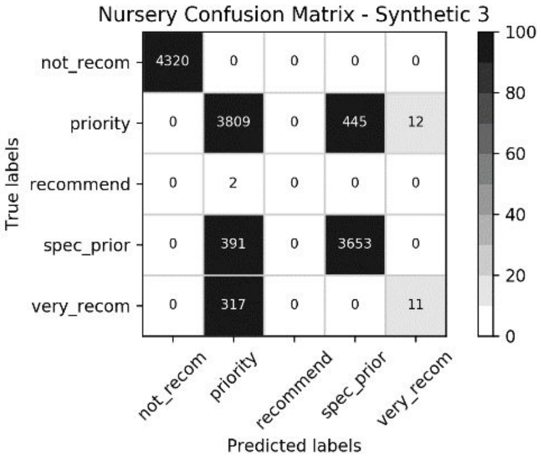
The confusion matrices for the performance of each of the five classifiers, trained on 100% of each of the eleven datasets (original, 5 synthetic non-parametric and 5 synthetic parametric) and tested on 100% of the original dataset are shown in Figure 3.2.11-3.2.15 for the Linear model, Decision Tree model, KNN model, Random Forest model and SVM model respectively. The results from cross comparison of the Nursery dataset when models are trained with synthetic data and tested with real data correlates with the earlier results where a higher degree of misclassification in the Nursery dataset, compared with the Breast Cancer dataset is observed. Again, this misclassification is observed in models trained with the real data and the synthetic data to a similar degree and so the problem may be the models used, and not the synthetic data.



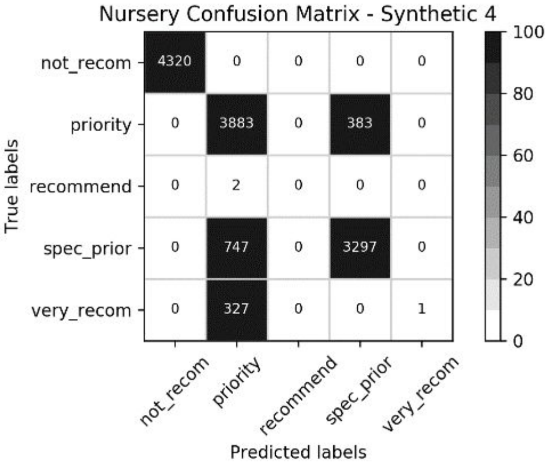




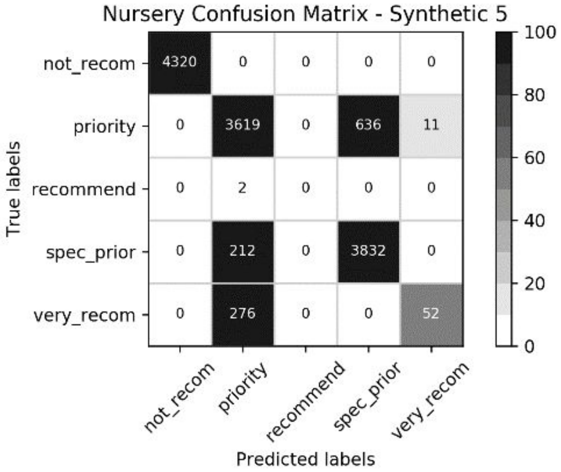
	not_recom	priority	recommend	spec_prior	very_recom
not_recom	4320	0	0	0	0
priority	1	3492	0	764	9
recommend	0	0	0	0	2
spec_prior	0	198	0	3846	0
very_recom	0	286	0	0	42



	not_recom	priority	recommend	spec_prior	very_recom
not_recom	4320	0	0	0	0
priority	0	3809	0	445	12
recommend	0	2	0	0	0
spec_prior	0	391	0	3653	0
very_recom	0	317	0	0	11

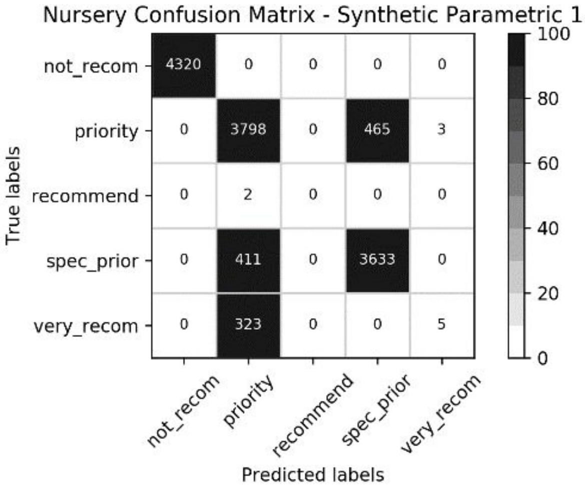


	not_recom	priority	recommend	spec_prior	very_recom
not_recom	4320	0	0	0	0
priority	0	3883	0	383	0
recommend	0	2	0	0	0
spec_prior	0	747	0	3297	0
very_recom	0	327	0	0	1



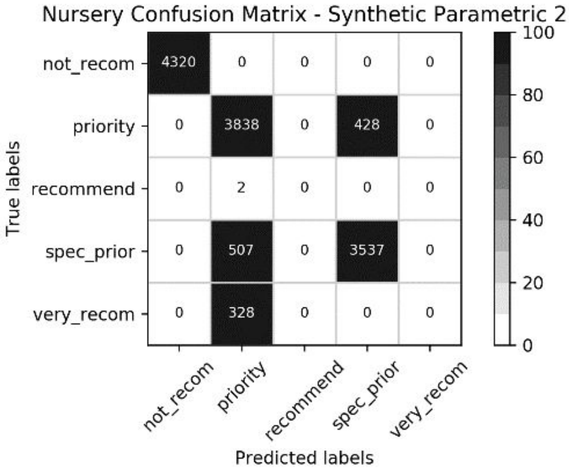
Nursery Confusion Matrix - Synthetic 5

True labels \ Predicted labels	not_recom	priority	recommend	spec_prior	very_recom
not_recom	4320	0	0	0	0
priority	0	3619	0	636	11
recommend	0	2	0	0	0
spec_prior	0	212	0	3832	0
very_recom	0	276	0	0	52



Nursery Confusion Matrix - Synthetic Parametric 1

True labels \ Predicted labels	not_recom	priority	recommend	spec_prior	very_recom
not_recom	4320	0	0	0	0
priority	0	3798	0	465	3
recommend	0	2	0	0	0
spec_prior	0	411	0	3633	0
very_recom	0	323	0	0	5



Nursery Confusion Matrix - Synthetic Parametric 2

True labels \ Predicted labels	not_recom	priority	recommend	spec_prior	very_recom
not_recom	4320	0	0	0	0
priority	0	3838	0	428	0
recommend	0	2	0	0	0
spec_prior	0	507	0	3537	0
very_recom	0	328	0	0	0

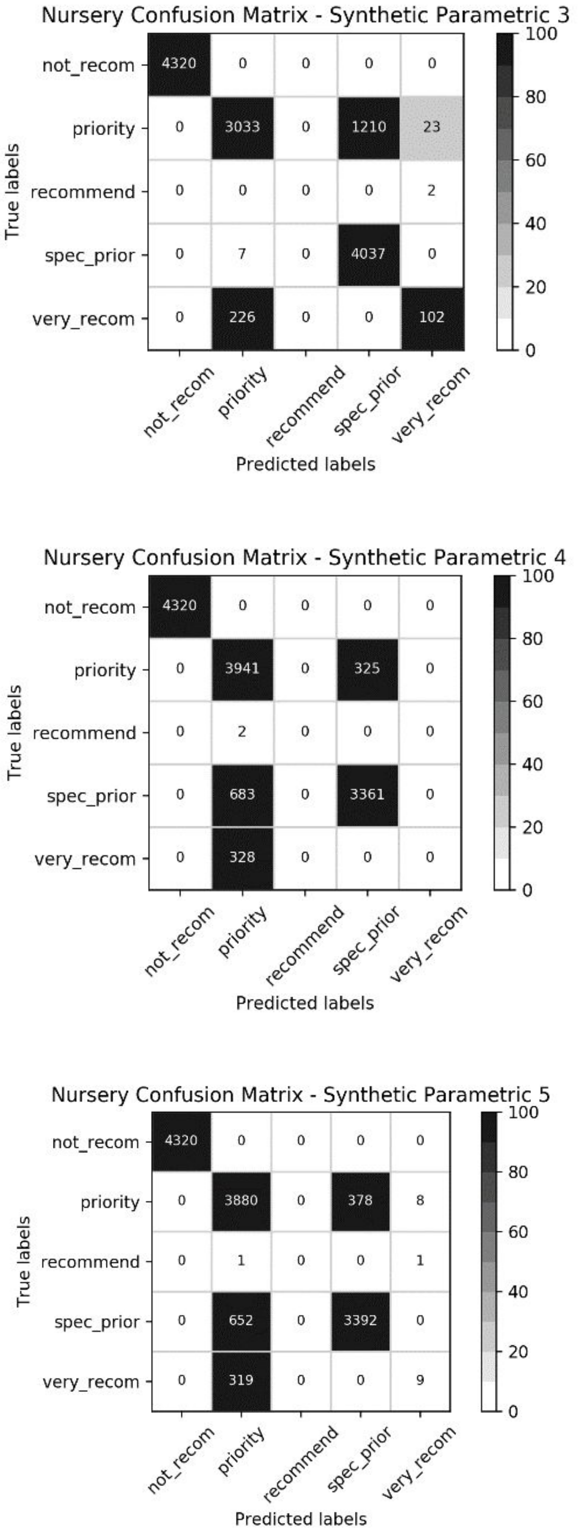
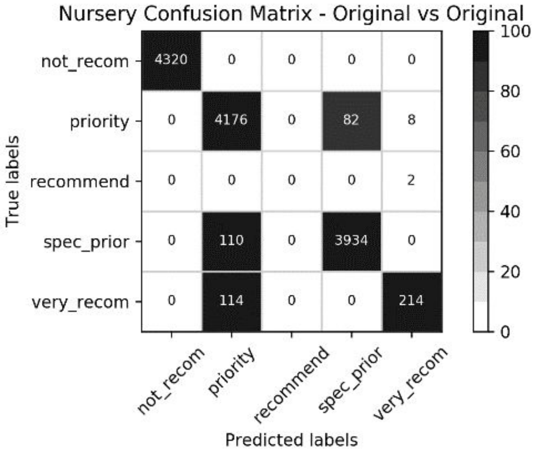
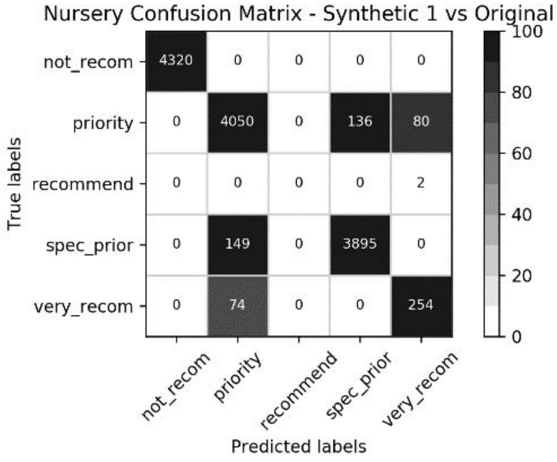
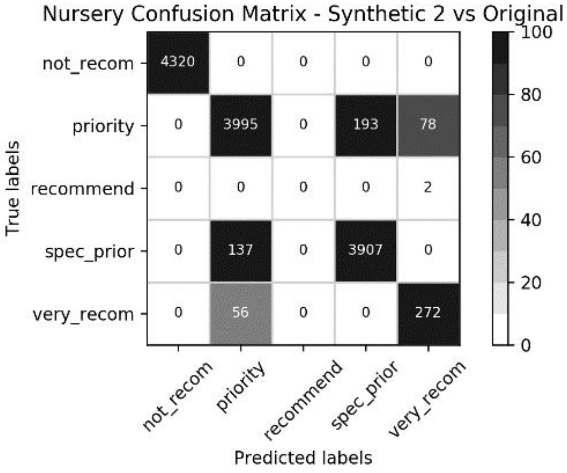
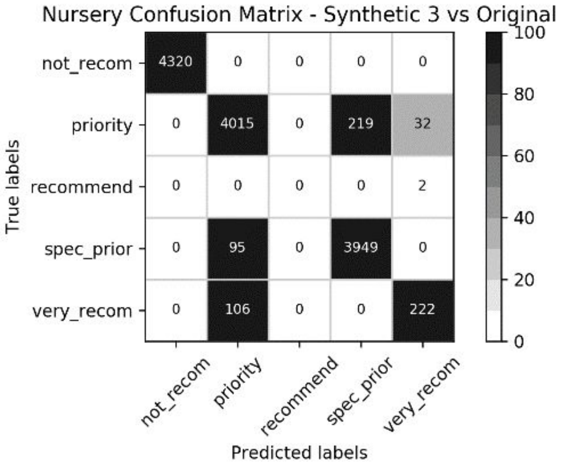


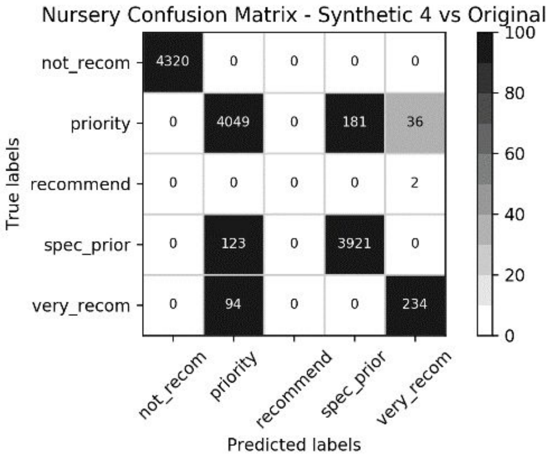
Figure 3.2.11 Confusion Matrices for the Linear Model when trained with each of the 11 datasets (1 original and 10 synthetic) and tested on 100% of the original dataset

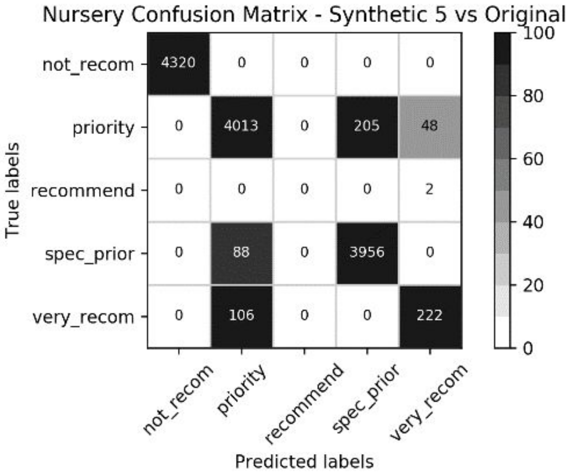

Nursery Confusion Matrix - Original vs Original
True labels
not_recom
priority
recommend
spec_prior
very_recom
4320 0 0 0 0
0 4176 0 82 8
0 0 0 0 2
0 110 0 3934 0
0 114 0 0 214
Predicted labels
not_recom
priority
recommend
spec_prior
very_recom
100
80
60
40
20
0

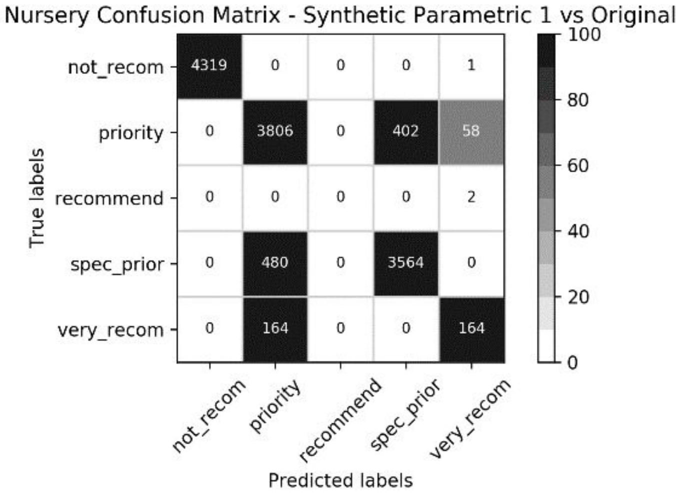

Nursery Confusion Matrix - Synthetic 1 vs Original
True labels
not_recom
priority
recommend
spec_prior
very_recom
4320 0 0 0 0
0 4050 0 136 80
0 0 0 0 2
0 149 0 3895 0
0 74 0 0 254
Predicted labels
not_recom
priority
recommend
spec_prior
very_recom
100
80
60
40
20
0


Nursery Confusion Matrix - Synthetic 2 vs Original
True labels
not_recom
priority
recommend
spec_prior
very_recom
4320 0 0 0 0
0 3995 0 193 78
0 0 0 0 2
0 137 0 3907 0
0 56 0 0 272
Predicted labels
not_recom
priority
recommend
spec_prior
very_recom
100
80
60
40
20
0

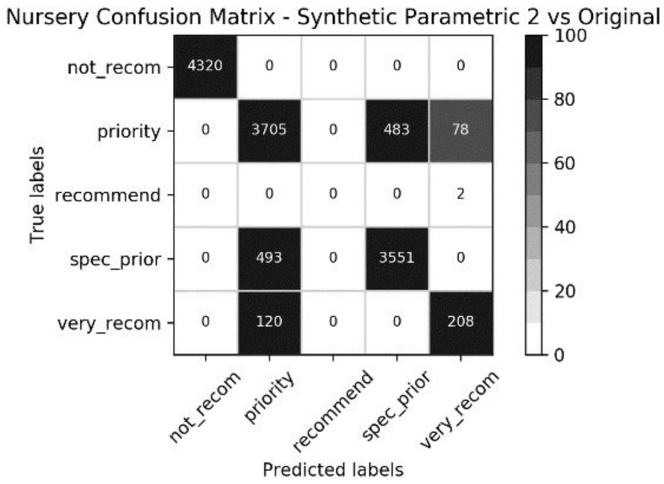

Nursery Confusion Matrix - Synthetic 3 vs Original


Nursery Confusion Matrix - Synthetic 4 vs Original

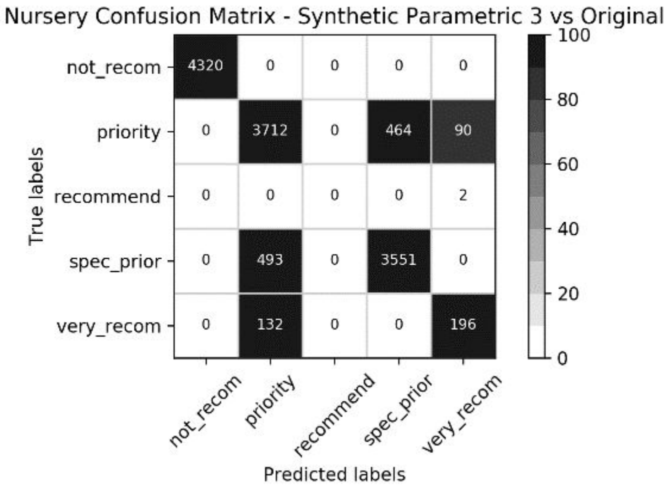

Nursery Confusion Matrix - Synthetic 5 vs Original



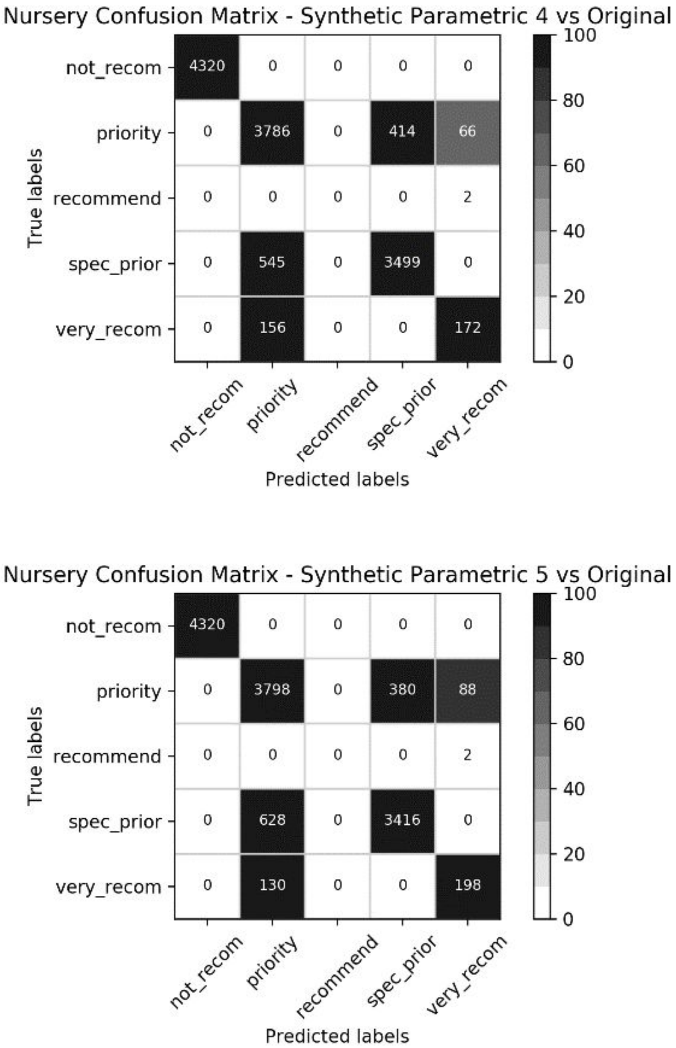
	not_recom	priority	recommend	spec_prior	very_recom
not_recom	4319	0	0	0	1
priority	0	3806	0	402	58
recommend	0	0	0	0	2
spec_prior	0	480	0	3564	0
very_recom	0	164	0	0	164



	not_recom	priority	recommend	spec_prior	very_recom
not_recom	4320	0	0	0	0
priority	0	3705	0	483	78
recommend	0	0	0	0	2
spec_prior	0	493	0	3551	0
very_recom	0	120	0	0	208



	not_recom	priority	recommend	spec_prior	very_recom
not_recom	4320	0	0	0	0
priority	0	3712	0	464	90
recommend	0	0	0	0	2
spec_prior	0	493	0	3551	0
very_recom	0	132	0	0	196



Nursery Confusion Matrix - Synthetic Parametric 4 vs Original

Nursery Confusion Matrix - Synthetic Parametric 5 vs Original

Figure 3.2.12 Confusion Matrices for the Decision Tree Model when trained with each of the 11 datasets (1 original and 10 synthetic) and tested on 100% of the original dataset

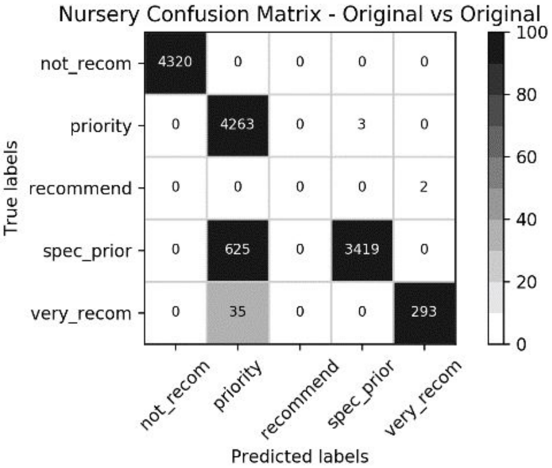

Nursery Confusion Matrix - Original vs Original
True labels
Predicted labels
not_recom
priority
recommend
spec_prior
very_recom
4320
0
0
0
0
0
4263
0
3
0
0
0
0
0
2
0
625
0
3419
0
0
35
0
0
293
100
80
60
40
20
0

Figure 3.2.13 Confusion Matrices for the KNN Model when trained with each of the 11 datasets (1 original and 10 synthetic) and tested on 100% of the original dataset

Figure 3.2.14 Confusion Matrices for the Random Forest Model when trained with each of the 11 datasets (1 original and 10 synthetic) and tested on 100% of the original dataset

Figure 3.2.15 Confusion Matrices for the SVM Model when trained with each of the 11 datasets (1 original and 10 synthetic) and tested on 100% of the original dataset

4 Conclusion

The results of this work have shown that synthesised numerical data very closely retains the same statistical properties as the real data. Non-parametric methods produce synthetic data that shares more similarities with the real data than data synthesised using parametric methods, although parametric methods still perform well.

Synthesised categorical data results in higher deviations from the real data for both parametric and non-parametric methods. Further investigation is required to determine the cause of such deviations. Additional categorical datasets will be synthesised and the results analysed in future work. Alternative encoding methods for categorical data will also be considered to determine if this has an impact on performance.

The performance of synthetic data was evaluated by creating classification models from both the real and synthetic data and comparing the performance of each, for both the numerical Breast Cancer dataset and the categorical Nursery dataset. Using 10-fold cross validation, models were trained and tested on each of the 11 datasets (1 real, 5 synthetic parametric and 5 synthetic non-parametric), for both the numerical and categorical datasets. The performance of the models trained and tested using synthetic data was compared with the performance of the model trained and tested using real data. In addition, for each dataset, and each of the 5 classifiers, models were created that were trained only on the synthetic data (1 model per synthetic dataset, for each classification algorithm). These models were then tested using all of the real data. The purpose was to determine whether synthetic data, generated from real data, was good enough to train models that could then be used in future to classify real observations correctly.

The differences observed in the accuracy of the models created with numerical data are negligible, for both parametric and non-parametric synthesising methods, with the non-parametric method achieving slightly better results than the non-parametric method. Models generated with real and synthetic data achieve high accuracy overall and therefore in this case, synthetic data could be considered a valid alternative to the real data.

Models generated using categorical data do not perform as well as those generated using numerical data. However, the results achieved are similar across the real and synthetic data. Therefore the performance issue may relate to the suitability of the selected machine learning algorithms used as opposed to the data itself.

False positives and false negatives were observed in the resulting confusion matrices from these experiments. These are relatively low for numerical data models but higher in categorical data models. This presents a problem when analysing data at the granular patient level, for example, when classifying whether a tumour is benign or malignant, a false negative can have very serious consequences. However, within the MIDAS project, data is being analysed at the population level for health care policy making. In this case small instances of false positives and false negatives may have a lesser impact on the results, as we are more interested in questions such as: “What region has the highest incidence of malignant tumours?” Population level analyses will be investigated further in the next iteration of this deliverable.

Whilst this work has shown that synthetic data, generated using the parametric and non-parametric methods in the SynthPop library, performs very similarly to real data when utilised in machine learning, further investigation is required with a broader range of datasets, numerical and categorical, and with more machine learning algorithms to provide a more rigorous and robust evaluation.

In addition, this work has not considered the synthesis of multiple linked datasets, e.g. for tables in a relational database. SynthPop does not currently support the synthesis of linked data tables unless the tables are joined into one combined data file. However, joining tables can cause the loss of identifier fields and sequences in the data. Future work will explore the consequences of joining such data for synthesis. Alternative approaches to synthetic data generation, and in particular the promising Synthetic Data Vault technique (Patki, Wedge, and Veeramachaneni, 2016), as well as deep learning methods such as Generative Adversarial Networks (GAN), for generating synthetic data will also potentially be investigated in future and the performance compared with the parametric and CART (non-parametric) methods analysed in this work. These methods can purportedly synthesise linked data accurately. The validity of this will be examined.

The experimental work has shown that it is possible to retain data utility using the synthesising methods under investigation. This is a small study, but it may indicate that the evaluation of models built using synthetic data are reflective of the results that would be achieved if real data had been used. It is pertinent that disclosure risk is also analysed in the next phase of this work to determine whether any risk remains to the disclosure of confidential data.

If further research supports this hypothesis, then data scientists could potentially mine synthetic healthcare datasets with an assumption that any knowledge elicited is very likely to be reflected in the real dataset at a population level. Using synthetic datasets to facilitate privacy preserving machine learning to discover patterns and enable viable predictive modelling without disclosing sensitive data has the potential to revolutionise health care research in an impactful way by opening up serious health care research that could drive improvements in population health and wellbeing much more quickly than is currently observed.

5 Dissemination of Synthetic Datasets

The original real datasets, as well as the synthetic datasets generated from this work for D3.11 are available in Dropbox. As the dissemination level of this deliverable is Public, this link is available for anyone to access. The datasets are available at the following link: https://bit.ly/2NsAmGi

6 References

Eno, J. and Thompson, C. (2008). Generating synthetic data to match data mining patterns. IEEE Internet Computing, 12(3), pp. 78-82.

Heyburn, R., Bond, R., Black, M., Mulvenna, M., Wallace, J., Rankin, D. and Cleland, B. (2018). Machine learning using synthetic and real data: Similarity of evaluation metrics for different healthcare datasets and for different algorithms. In: Data Science and Knowledge Engineering for Sensing Decision Support, pp.1281-1291.

IBM Corporation (2017). 10 Key Marketing Trends for 2017. [online] IBM Marketing Cloud. Available at: http://whitepapers.insidebigdata.com/content62746 [Accessed 1 Jun. 2019].

Little, R. (1993). Statistical Analysis of Masked Data. Journal of Official Statistics, 9, pp.407-426.

Mangasarian, O. and Wolberg, W. (1990). Cancer diagnosis via linear programming. SIAM News, 23(5), pp. 1 & 18.

Mangasarian, O., Setiono, R. and Wolberg, W. (1990). Pattern recognition via linear programming: Theory and application to medical diagnosis. In: Large-scale numerical optimization, SIAM Publications, Philadelphia, pp. 22-30.

Olave, M., Rajkovic, V., and Bohanec, M. (1989). An application for admission in public school systems. Expert Systems in Public Administration, pp. 145-160.

Nowok, B., Raab, G. and Dibben, C. (2016). synthpop: Bespoke Creation of Synthetic Data in R. Journal of Statistical Software, 74(11).

Patki, N., Wedge, R., and Veeramachaneni, K. (2016). The Synthetic Data Vault. In:IEEE International Conference on Data Science and Advanced Analytics (DSAA).

Raghunathan, T., Reiter, J. and Rubin, D. (2003). Multiple Imputation for Statistical Disclosure Limitation. Journal of Official Statistics, 19, pp.1-16.

Reiter, J. (2004a). New Approaches to Data Dissemination: A Glimpse into the Future (?). CHANCE, 17(3), pp.11-15.

Reiter, J. (2004b). Simultaneous use of multiple imputation for missing data and disclosure limitation. Survey Methodology, 30, pp.235-242.

Reiter, J. (2005a). Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. Journal of the Royal Statistical Society, Series A, 168(1), pp.185-205.

Reiter, J. (2005b). Significance tests for multi-component estimands from multiply imputed, synthetic microdata. Journal of Statistical Planning and Inference, 131(2), pp.365-377.

Reiter, J. (2005c). Using CART to generate partially synthetic public use microdata. Journal of Official Statistics, 21, pp. 441 - 462.

Reiter, J. and Raghunathan, T. (2007). The Multiple Adaptations of Multiple Imputation. Journal of the American Statistical Association, 102(480), pp.1462-1471.

Reiter, J. (2009). Using Multiple Imputation to Integrate and Disseminate Confidential Microdata. International Statistical Review, 77(2), pp.179-195.

Reiter, J. and Drechsler, J. (2010). Two stage multiple imputation to protect confidentiality. Statistica Sinica, 20, pp. 405 - 422.

Rubin, D. (1993). Statistical Disclosure Limitation. Journal of Official Statistics, 9(2), pp.461-468.

Wolberg, W. and Mangasarian, O. (1990). Multisurface method of pattern separation for medical diagnosis applied to breast cytology. In: Proceedings of the National Academy of Sciences, U.S.A., 87, pp. 9193-9196.

Zupan, M., Bohanec, M., Bratko, I., and Demsar, J. (1997). Machine learning by function decomposition. In: ICML-97, Nashville, TN.

7 Appendix A - Breast Cancer Dataset Decision Trees

Decision Tree - Original

Decision Tree - Non-Parametric V1

Decision Tree - Non-Parametric V2

Decision Tree - Non-Parametric V3

Decision Tree - Non-Parametric V4

Decision Tree - Non-Parametric V5

Decision Tree - Parametric V1

Decision Tree - Parametric V2

Decision Tree - Parametric V3

Decision Tree - Parametric V4

Decision Tree - Parametric V5

8 Appendix B - Breast Cancer Dataset Decision Trees Cross Comparison

Decision Tree - Original

Decision Tree - Non-Parametric V1

Decision Tree - Non-Parametric V2

Decision Tree - Non-Parametric V3

Decision Tree - Non-Parametric V4

Decision Tree - Non-Parametric V5

Decision Tree - Parametric V1

Decision Tree - Parametric V2

Decision Tree - Parametric V3

Decision Tree - Parametric V4

Decision Tree - Parametric V5