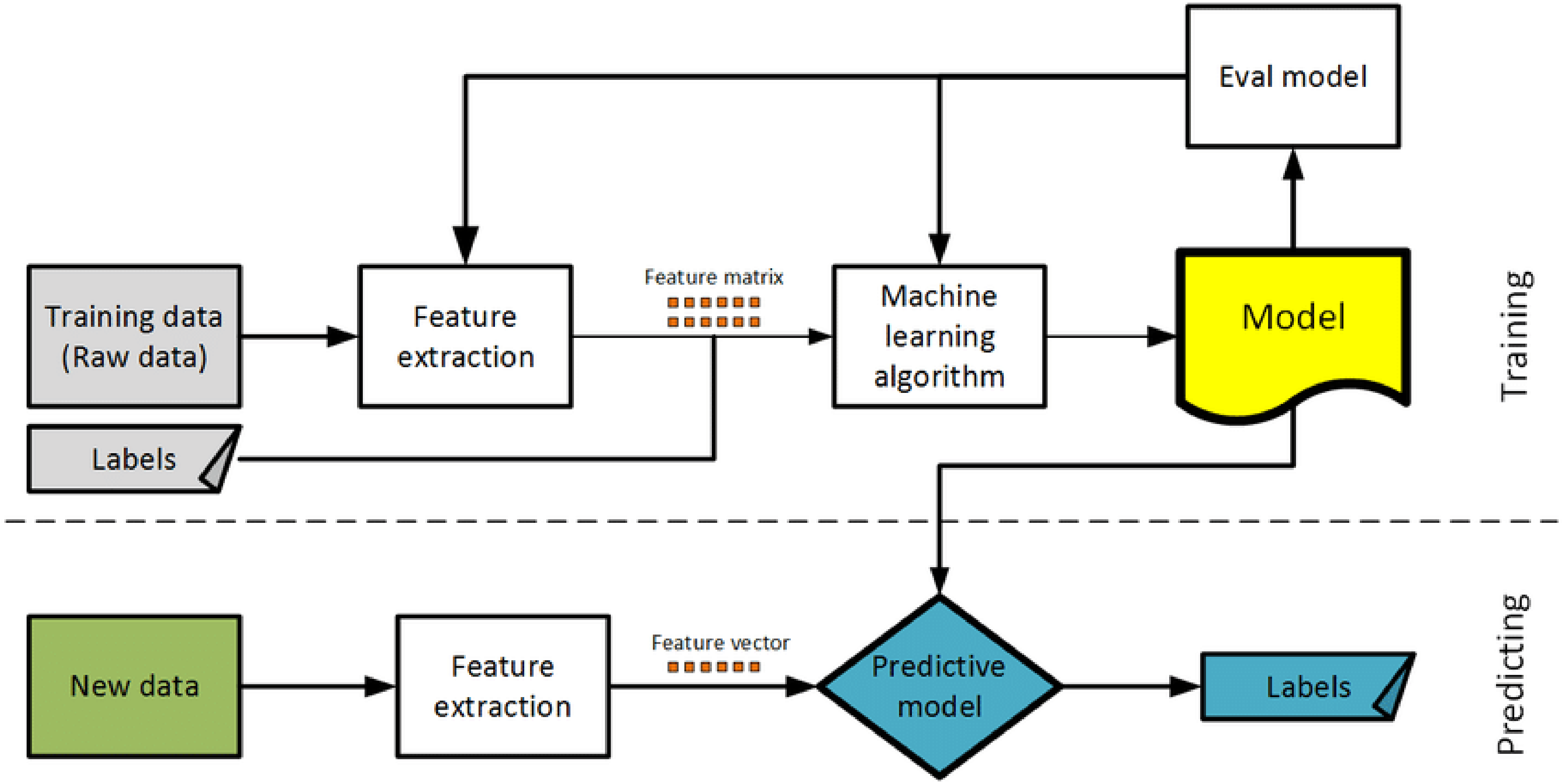
电子科技大学
格拉斯哥学院
Glasgow College, UESTC

Supervised Learning Workflow

State-of-the-Art



Source: https://www.researchgate.net/figure/A-flowchart-of-a-supervised-machine-learning-model_fig1_314202159

Potential Tasks

- **Data Preparation/Cleaning**: Identifying and correcting mistakes or errors in the data.

- **Feature Extraction/Selection**: Identifying those input variables that are most relevant to the task.

- **Data Transforms**: Changing the scale or distribution of variables.

- **Feature Engineering**: Deriving new variables from available data.

- **Dimensionality Reduction**: Creating compact projections of the data.

What is Data Preparation?

- Data preparation is the process of cleaning and transforming raw data prior to processing and analysis.
- It is an important step prior to processing and often involves reformatting data, making corrections to data, and combining datasets to enrich data.
- Data preparation is often a lengthy undertaking for data engineers or business users, but it is essential as a prerequisite to put data in context to turn it into insights and eliminate bias resulting from poor data quality.

Significance of Data Preparation

- Machine Learning algorithms are mathematical algorithms that use arithmetic operations to create prediction systems. Recall that these operations can only be performed on numbers, therefore suggesting that your dataset should be only numbered before it is sent as an input to the algorithms.

- Adding irrelevant features can deteriorate your model's performance. Besides that, choosing a subset of features also helps in saving on computation costs and time consumed because of model complexity.

- Certain machine learning models give the best results when their input is well-calibrated as per model structure.

What is Data Cleaning?

- Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset.
- When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled.
- If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct.
- There is no one absolute way to prescribe the exact steps in the data cleaning process because the processes will vary from dataset to dataset. But it is crucial to establish a template for your data cleaning process, so you know you are doing it the right way every time.

Handle Missing Data

- Missing data is a deceptively tricky issue in applied machine learning.
- Unfortunately, the two most commonly recommended methods of dealing with missing data are actually very bad.
- Dropping observations that have missing values
- Imputing the missing values based on other observations

The key is to tell your algorithm that the value was originally missing.

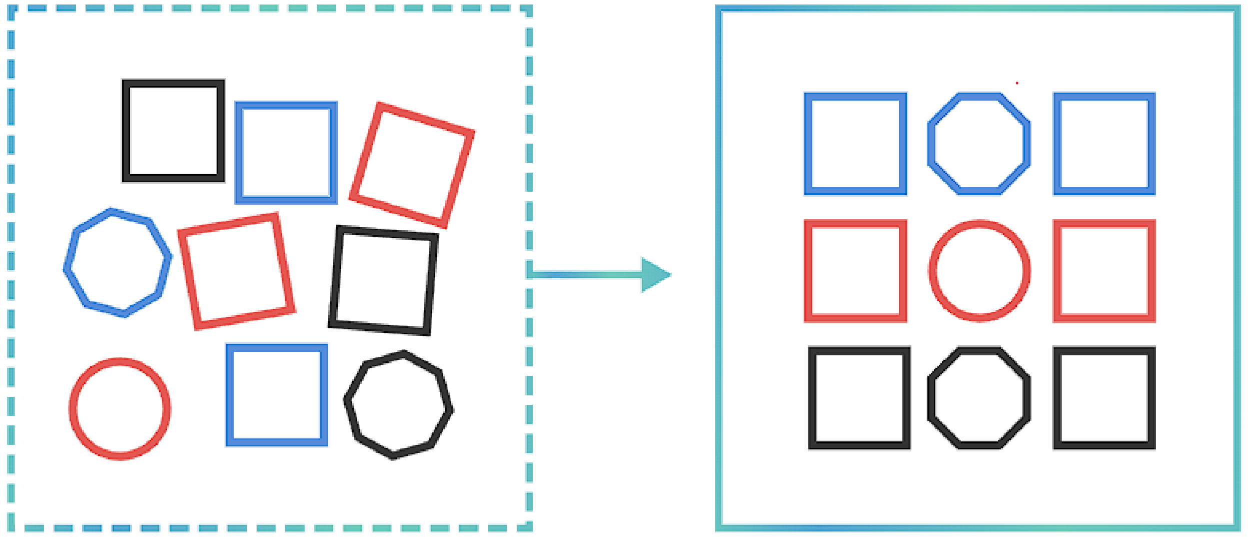
Handle Missing Data

- Dropping missing values is sub-optimal because when you drop observations, you drop **information**.
- The fact that the value was missing may be informative in itself. Plus, in the real world, you often need to make predictions on new data, even if some of the features are missing!

- Imputing missing values is sub-optimal because the value was originally missing, but you filled it in.
- This also leads to a loss of information, no matter how sophisticated your imputation method is. Even if you build an imputation model, you're just reinforcing the patterns already provided by other features.

What is Data Transforms?

Data transformation is the process of converting, cleansing, and structuring data into a usable format that can be analysed to support decision making processes.



Why Data Transforms?

- Transformation is an essential step in many processes, such as data integration, migration, warehousing and wrangling.

- The process of data transformation can be:
  - Constructive, where data is added, copied or replicated
  - Destructive, where records and fields are deleted
  - Aesthetic, where certain values are standardized, or
  - Structural, which includes columns being renamed, moved and combined

Data Transformation Techniques

Revising

- Revising ensures the data supports its intended use by organizing it in the required and correct way. It does this in a range of ways.

Manipulation

- This involves creation of new values from existing ones or changing current data through computation.
- Manipulation is also used to convert unstructured data into structured data that can be used by machine learning algorithms.

Separating

- This involves dividing up the data values into its parts for granular analysis. Splitting involves dividing up a single column with several values into separate columns with each of those values. This allows for filtering on the basis of certain values.

Data Transformation Techniques

Combining/ Integrating

Records from across tables and sources are combined to acquire a more holistic view of activities and functions. It couples data from multiple tables and datasets and combines records from multiple tables.

Data Smoothing

This process removes meaningless, noisy, or distorted data from the data set. By removing outliers and trends are most easily identified.

Data Aggregation

This technique gathers raw data from multiple sources and turns it into a summary form which can be used for analysis. An example is the raw data providing statistics such as averages and sums.

Data cleaning vs data transformation

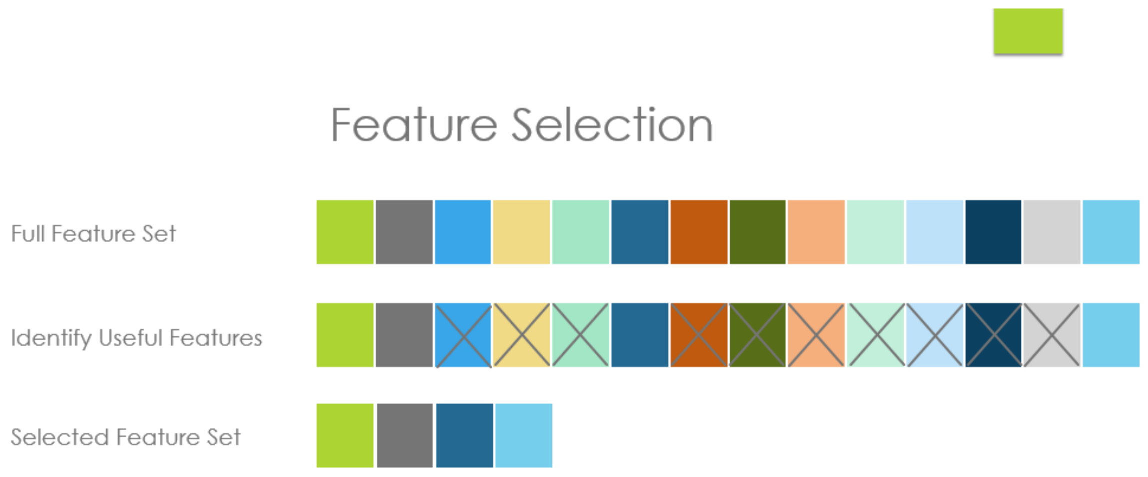
- Data cleaning is the process that removes data that does not belong in your dataset.
- Data transformation is the process of converting data from one format or structure into another.
- Transformation processes can also be referred to as data wrangling, transforming and mapping data from one "raw" data form into another format for warehousing and analysing.

What is Feature Selection?

Feature: A feature of something is an interesting or important part or characteristic of it.

Feature Selection is the method of reducing the input variable to your model by using only relevant data and getting rid of noise in data.

It is the process of automatically choosing relevant features for your machine learning model based on the type of problem you are trying to solve.



Why Feature Selection?

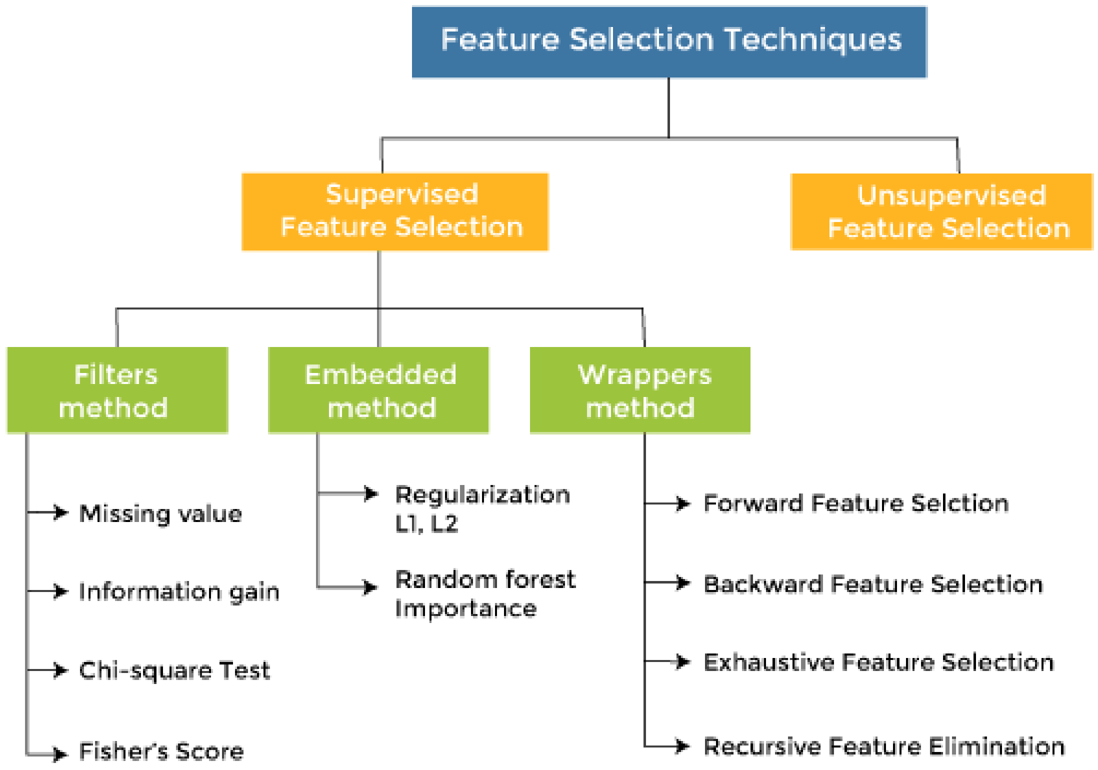
- Reduces Overfitting: Less redundant data means less opportunity to make decisions based on noise.
- Improves Accuracy: Less misleading data means modeling accuracy improves.
- Reduces Training Time: fewer data points reduce algorithm complexity and algorithms train faster.

Feature Selection Techniques

There are mainly two types of Feature Selection techniques, which are:

- Unsupervised Feature selection techniques ignore the target variable and can be used for the unlabelled dataset.

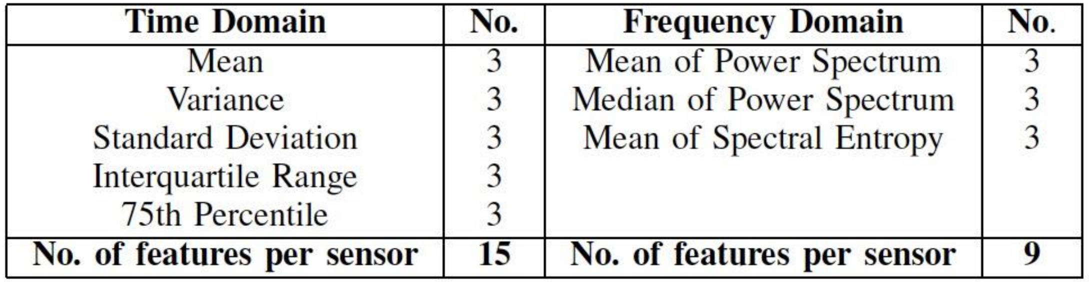
- Supervised Feature selection techniques consider the target variable and can be used for the labelled dataset.



An Example from our Study

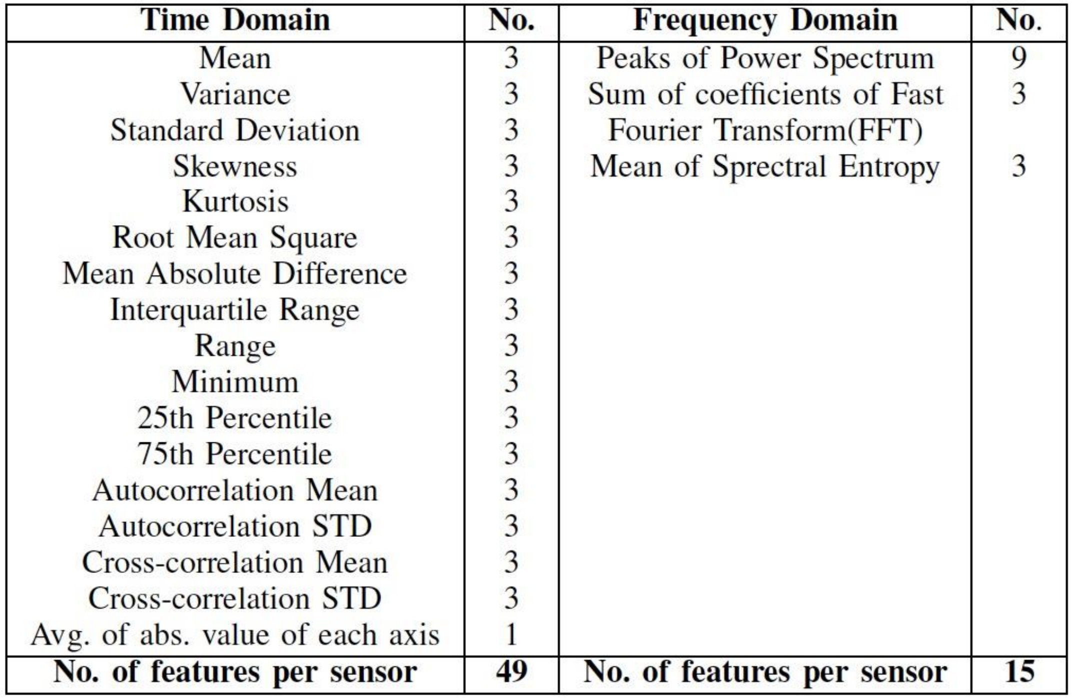
- Table II: A total of 64 features from one accelerometer, or 192 from all 3.
- Table III: A total of 24 features from one accelerometer, or 72 from all 3.
- The SVC obtained an overall accuracy of 75.3% on the entire 192 features of the first feature set, and 76.2% on the 72 features of the second.

TABLE III
FEATURE SET 2



Time Domain	No.	Frequency Domain	No.
Mean	3	Mean of Power Spectrum	3
Variance	3	Median of Power Spectrum	3
Standard Deviation	3	Mean of Spectral Entropy	3
Interquartile Range	3		
75th Percentile	3		
No. of features per sensor	**15**	**No. of features per sensor**	**9**

TABLE II
FEATURE SET 1



Time Domain	No.	Frequency Domain	No.
Mean	3	Peaks of Power Spectrum	9
Variance	3	Sum of coefficients of Fast Fourier Transform(FFT)	3
Standard Deviation	3		
Skewness	3	Mean of Sprectral Entropy	3
Kurtosis	3		
Root Mean Square	3		
Mean Absolute Difference	3		
Interquartile Range	3		
Range	3		
Minimum	3		
25th Percentile	3		
75th Percentile	3		
Autocorrelation Mean	3		
Autocorrelation STD	3		
Cross-correlation Mean	3		
Cross-correlation STD	3		
Avg. of abs. value of each axis	1		
No. of features per sensor	**49**	**No. of features per sensor**	**15**

An Example from our Study

- Table II: For the first set, forward sequential feature selection (SFS) selected just 18 features and obtained an accuracy of **83.4%**, while backward SFS selected 174 features with an accuracy of only 76.4%. For the second set,

- Table III: Both forward and backward SFS yielded similar results, with the former having selected 18 features and the latter, 20, with accuracies of **82.2%** for each, respectively.

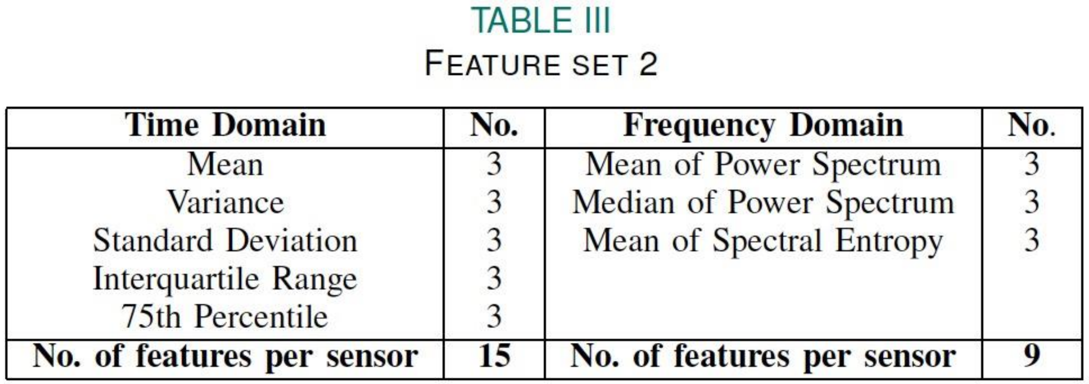


TABLE III
FEATURE SET 2

Time Domain	No.	Frequency Domain	No.
Mean	3	Mean of Power Spectrum	3
Variance	3	Median of Power Spectrum	3
Standard Deviation	3	Mean of Spectral Entropy	3
Interquartile Range	3		
75th Percentile	3		
No. of features per sensor	**15**	**No. of features per sensor**	**9**

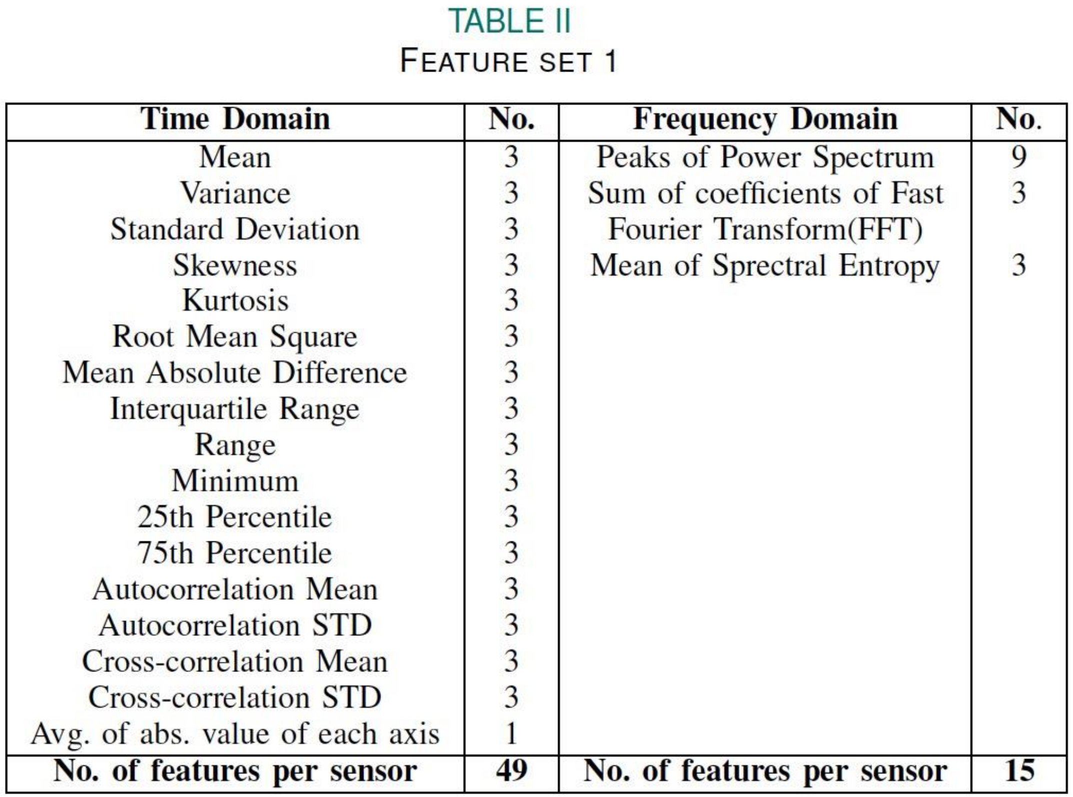
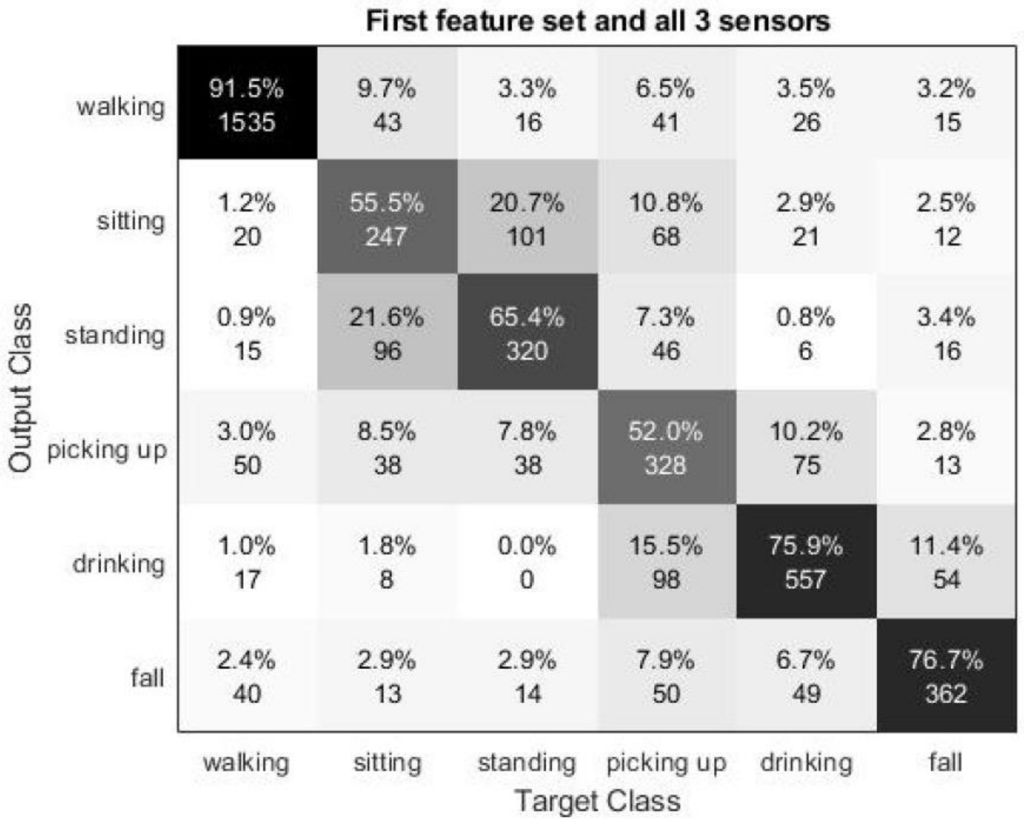


TABLE II
FEATURE SET 1

Time Domain	No.	Frequency Domain	No.
Mean	3	Peaks of Power Spectrum	9
Variance	3	Sum of coefficients of Fast Fourier Transform(FFT)	3
Standard Deviation	3		
Skewness	3	Mean of Sprectral Entropy	3
Kurtosis	3		
Root Mean Square	3		
Mean Absolute Difference	3		
Interquartile Range	3		
Range	3		
Minimum	3		
25th Percentile	3		
75th Percentile	3		
Autocorrelation Mean	3		
Autocorrelation STD	3		
Cross-correlation Mean	3		
Cross-correlation STD	3		
Avg. of abs. value of each axis	1		
No. of features per sensor	**49**	**No. of features per sensor**	**15**

An Example from our Study

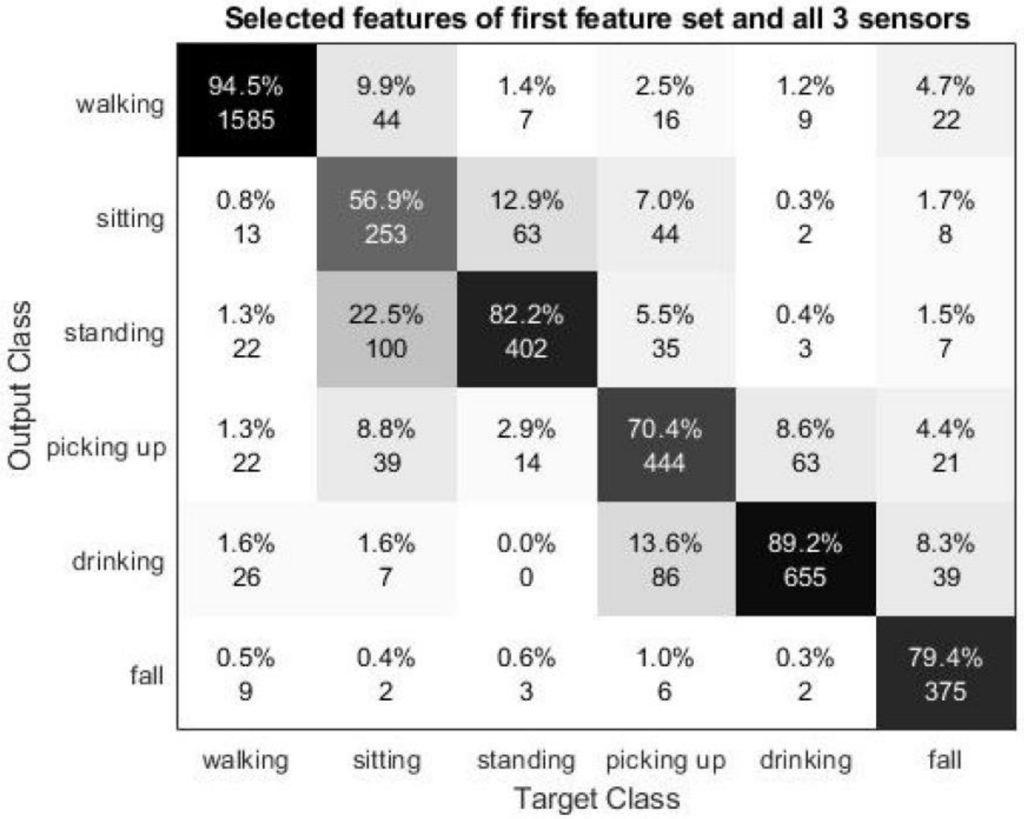


First feature set and all 3 sensors

Output Class
Target Class

	walking	sitting	standing	picking up	drinking	fall
walking	91.5% 1535	9.7% 43	3.3% 16	6.5% 41	3.5% 26	3.2% 15
sitting	1.2% 20	55.5% 247	20.7% 101	10.8% 68	2.9% 21	2.5% 12
standing	0.9% 15	21.6% 96	65.4% 320	7.3% 46	0.8% 6	3.4% 16
picking up	3.0% 50	8.5% 38	7.8% 38	52.0% 328	10.2% 75	2.8% 13
drinking	1.0% 17	1.8% 8	0.0% 0	15.5% 98	75.9% 557	11.4% 54
fall	2.4% 40	2.9% 13	2.9% 14	7.9% 50	6.7% 49	76.7% 362

Fig. 6. Results using all 192 features of set 1



Selected features of first feature set and all 3 sensors

Output Class
Target Class

	walking	sitting	standing	picking up	drinking	fall
walking	94.5% 1585	9.9% 44	1.4% 7	2.5% 16	1.2% 9	4.7% 22
sitting	0.8% 13	56.9% 253	12.9% 63	7.0% 44	0.3% 2	1.7% 8
standing	1.3% 22	22.5% 100	82.2% 402	5.5% 35	0.4% 3	1.5% 7
picking up	1.3% 22	8.8% 39	2.9% 14	70.4% 444	8.6% 63	4.4% 21
drinking	1.6% 26	1.6% 7	0.0% 0	13.6% 86	89.2% 655	8.3% 39
fall	0.5% 9	0.4% 2	0.6% 3	1.0% 6	0.3% 2	79.4% 375

Fig. 7. Results using all 18 selected features of set 1

What is Feature Engineering?

- Feature engineering is a technique that leverages data to create new variables that aren't in the training set.

- It can produce new features for both supervised and unsupervised learning, with the goal of simplifying and speeding up data transformations while also enhancing model accuracy.

What is Feature Engineering?

- Feature engineering in Machine Learning involves extracting useful features from given input data following the target to be learned and the machine learning model used.
- It involves transforming data to forms that better relate to the underlying target to be learned.
- When done right, feature engineering can augment the value of your existing data and improve the performance of your machine learning models.
- On the other hand, using bad features may require you to build much more complex models to achieve the same level of performance.

Feature Engineering Techniques

Handling Outliers: The various methods of handling outliers include:

- Removal: Outlier-containing entries are deleted from the distribution.
- Replacing values: Alternatively, the outliers could be handled as missing values and replaced with suitable imputation.
- Capping: Using an arbitrary value or a value from a variable distribution to replace the maximum and minimum values.
- Discretization: Discretization is the process of converting continuous variables, models, and functions into discrete ones.

Feature Engineering Techniques

Log Transform:

Log Transform is the most used technique among data scientists. It’s mostly used to turn a skewed distribution into a normal or less-skewed distribution.

$$data = \log(\mathrm{data})$$

Scaling:

Normalization: All values are scaled in a specified range between 0 and 1 via normalisation (or min-max normalisation).

Standardization: Standardization (also known as z-score normalisation) is the process of scaling values while accounting for standard deviation.

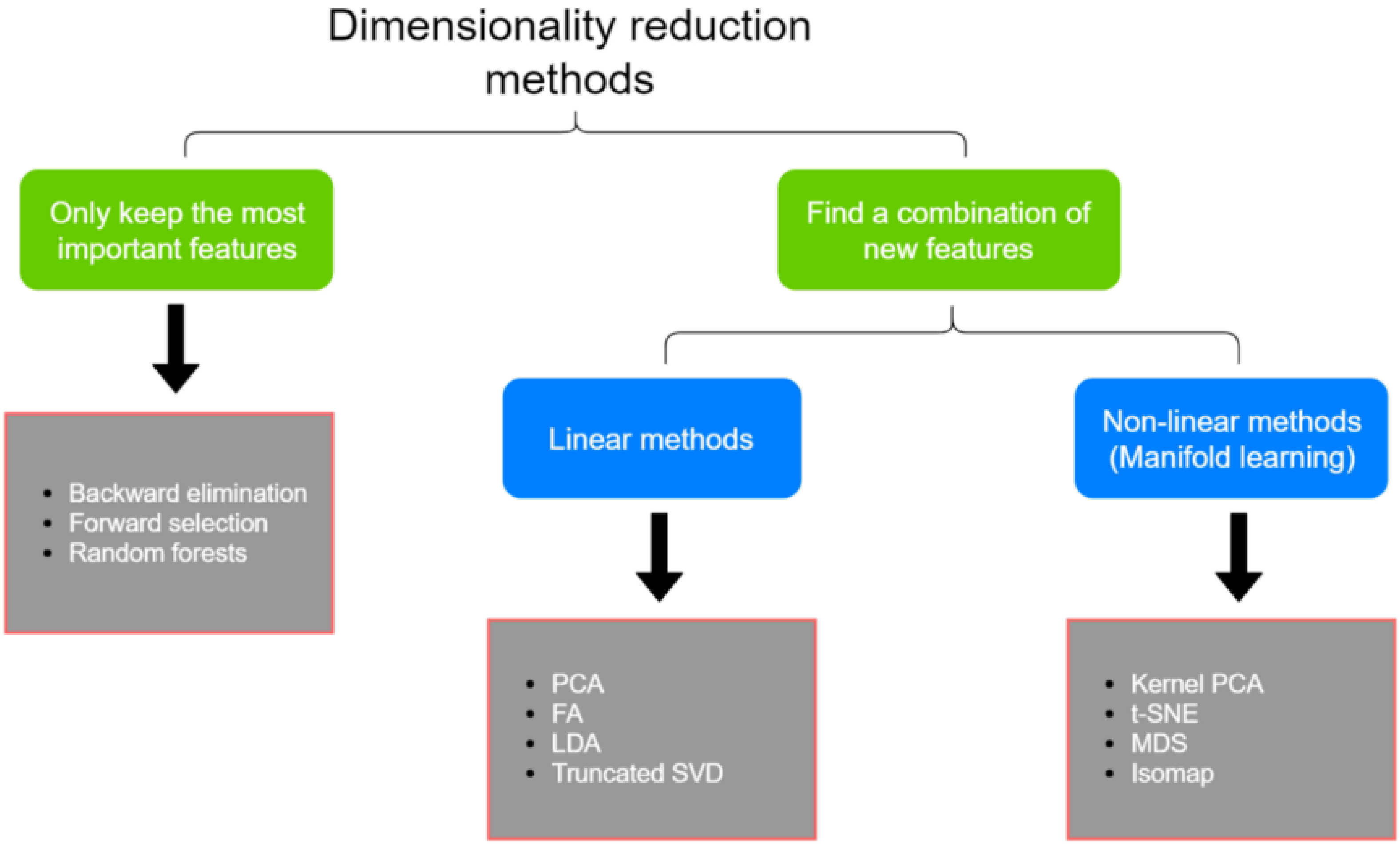
Problem With Many Input Variables

- The performance of machine learning algorithms can degrade with too many input variables.

- Having a large number of dimensions in the feature space can mean that the volume of that space is very large, and in turn, the points that we have in that space (rows of data) often represent a small and non-representative sample.

- This can dramatically impact the performance of machine learning algorithms fit on data with many input features, generally referred to as the “curse of dimensionality.”

Dimensionality Reduction Methods



Dimensionality reduction methods

Only keep the most important features

- Backward elimination
- Forward selection
- Random forests

Find a combination of new features

Linear methods

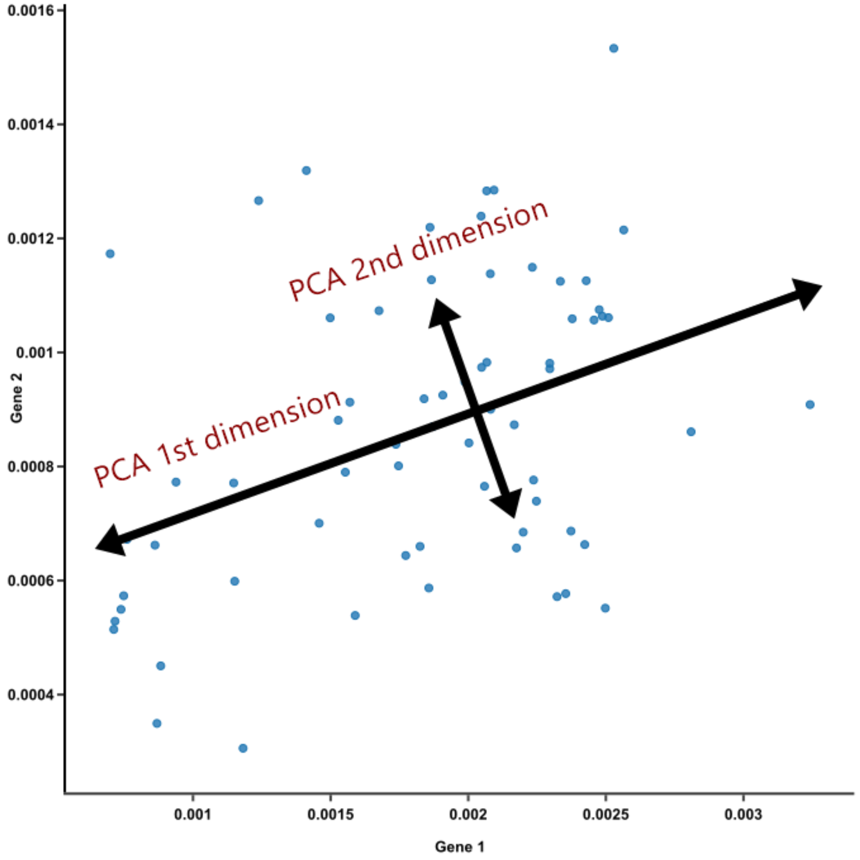
- PCA
- FA
- LDA
- Truncated SVD

Non-linear methods (Manifold learning)

- Kernel PCA
- t-SNE
- MDS
- Isomap

Principal Component Analysis (PCA)

PCA is a linear dimensionality reduction technique that transforms a set of correlated variables (*p*) into a smaller number of uncorrelated variables called Principal Components while retaining as much of the variation in the original dataset as possible.

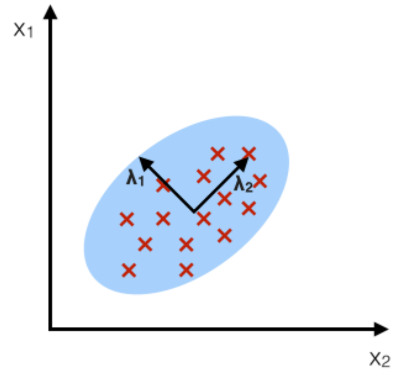


Linear Discriminant Analysis (LDA)

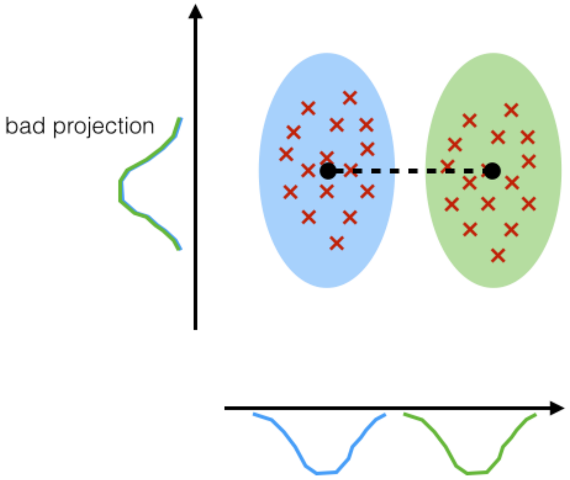
LDA is typically used for multi-class classification. It can also be used as a dimensionality reduction technique.

LDA best separates or discriminates (hence the name LDA) training instances by their classes.

PCA:
component axes that maximize the variance



LDA:
maximizing the component axes for class-separation



bad projection