# An Event Group Based Classification Framework for Multi-variate Sequential Data

**Chao Sun**
The University of Sydney
chao.sun@sydney.edu.au

**David Stirling**
University of Wollongong

## Abstract

Decision tree algorithms were not traditionally considered for sequential data classification, mostly because feature generation needs to be integrated with the modelling procedure in order to avoid a localisation problem. This paper presents an Event Group Based Classification (EGBC) framework that utilises an X-of-N (XoN) decision tree algorithm to avoid the feature generation issue during the classification on sequential data. In this method, features are generated independently based on the characteristics of the sequential data. Subsequently an XoN decision tree is utilised to select and aggregate useful features from various temporal and other dimensions (as event groups) for optimised classification. This leads the EGBC framework to be adaptive to sequential data of differing dimensions, robust to missing data and accommodating to either numeric or nominal data types. The comparatively improved outcomes from applying this method are demonstrated on two distinct areas – a text based language identification task, as well as a honeybee dance behaviour classification problem. A further motivating industrial problem – hot metal temperature prediction, is further considered with the EGBC framework in order to address significant real-world demands.

**Keywords**: Multi-variate time series; Symbolic data mining; Pattern search; SAX motifs; X-of-N decision trees

## 1 Introduction

Time series data mining (TSDM) is a challenging task, which has attracted enormous attention in the recent years. TSDM is of particular interest in many industry areas where improvements and knowledge extensions are expected from utilising data science technology over massive existing operational records. The research work described in this paper was initiated and motivated by the real industrial needs from the iron-making industry, where operational data was continuously collected from a Blast Furnace (BF) for a number of years. Through building predictive models on the BF data, the authors have identified a major problem that the data attributes, due to the variation of type or the spatial difference at collection, are shifted along the time axis with different extent. As the results may occur and be detect prior to the detection of the causes, the inter-variable Granger causality relationship (Granger 1969) are broken.

The situation in which attributes are not aligned in time is in fact a quite common problem in TSDM, however this is often overlooked in practice. Temporal misalignment is the name we attribute to this scenario where time delays exist between the generation and detection of various data attribution, and therefore break the causal-resultant relationship between input and output attributes. The time shifts on differing attributes are generally not uniform; these in turn make synchronisation very hard if not possible. For example, Figure 1 illustrates a typical example of temporal misalignment when lightning and thunder are detected at different times with different time spans due to the speed difference between light and sound.

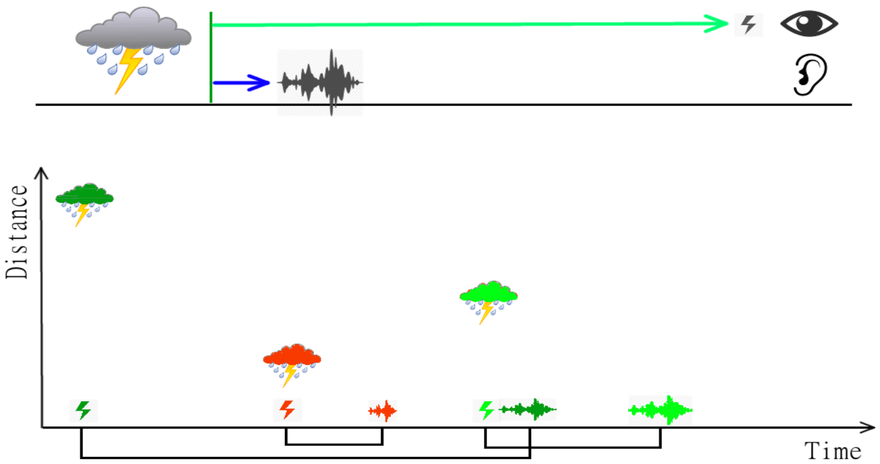


*Figure 1: Temporal misalignment in the lightning detection. The heights of the cloud signs (on Y-Axis) represent the distances between the lightning events and the observer. And the X-Axis represents the time when flashes and the thunder are detected. Assuming the flashes are detected almost immediately, the delays of the associated thunder (marked with the same colour) are determined by the associated distance. Therefore, proximally close flash and thunder pairs may in-fact not be associated to each other, due to the occurrence of temporal misalignment.*

In this example, when multiple lightning events (marked as dark/light green and red clouds) occur at different distances (Y-Axis) from the observer, the corresponding flashes and thunder (marked by the same colours as the source clouds) may be received in a mixed ordering that are different than the ordering of the source lightning. Without prior knowledge of the distance of each lightning event, the time shift as a function of distance cannot be recovered, therefore it is impossible to correctly associate the corresponding flash and thunder observations.

The lightning example illustrates that even in such a simple system with only two attributes, without some awareness of additional contextual information, such as the distance, height and direction between the actual lightning and the observer, correctly synchronising the light and sound signals is very difficult. Due to the complex and stochastic nature of the industrial processes in the BF, it is near impossible to analyse this huge industrial plant using the fixed attribute-value model data mining techniques. This paper proposes a new approach, the Event Group Based Classification (EGBC) framework in order to address this temporal misalignment issue.

In the remainder of this paper, Section 2 provides background and relevant research work. Section 3 focuses on the methodology and experimental settings. Section 4 describes a number of experimental results by applying the EGBC framework to various time series classification tasks and compares the outcomes with other existing techniques. Lastly, in Section 5 we discuss the advantages of this approach and the possibility to further develop this as a universal, time series data mining framework for solving other real-world problems.

## 2 Background and Literature Review

A BF, as illustrated in Figure 2, is a massive industrial plant that functions as both a chemical reactor and a heat exchanger continuously for many years. Iron oxide pellets, or sintered ore, along with coke and limestone, are charged at the top into layers, which gradually over some eight hours decent in the blast furnace undergoing a series of chemical reductions to produce molten iron.

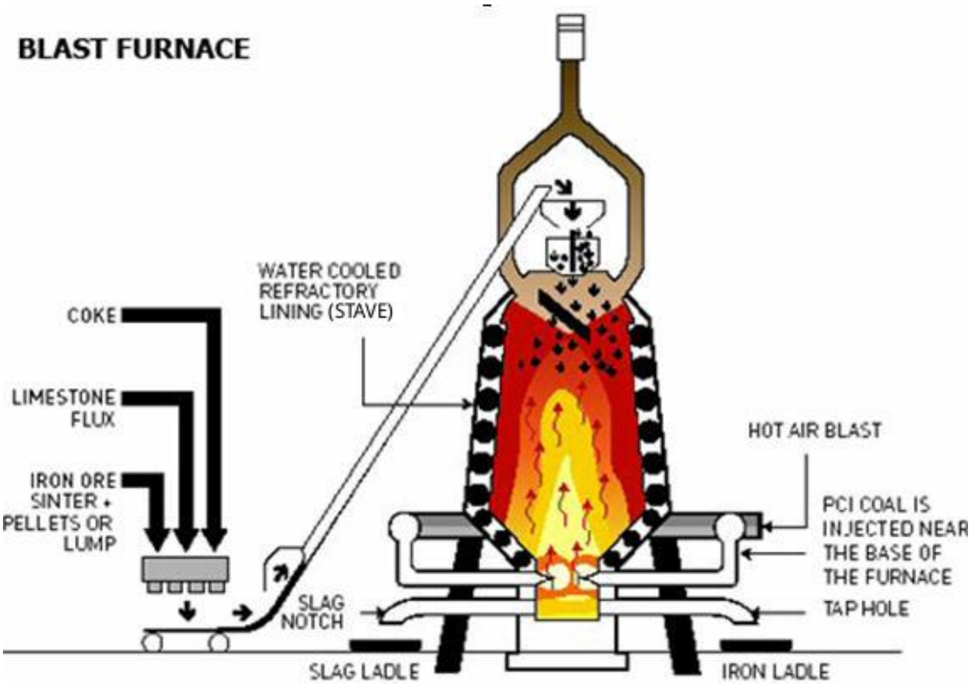

BLAST FURNACE
COKE
LIMESTONE FLUX
IRON ORE SINTER + PELLETS OR LUMP
WATER COOLED REFRACTORY LINING (STAVE)
HOT AIR BLAST
PCI COAL IS INJECTED NEAR THE BASE OF THE FURNACE
TAP HOLE
SLAG NOTCH
SLAG LADLE
IRON LADLE

*Figure 2: A diagram of the blast furnace in the iron-making industry.*

The BF is a typical non-stationary system; almost all attributes in the BF dataset are subject to uncertain time shifts. For instance, domain experts explain that newly produced molten iron may require one to two hours before it is sufficiently tapped out of the BF. Other production related data, such as the contents of the exhaust gases, metallurgy lab test results and the quantity of materials fed from the top of the BF, are generally acquired at a variety of sampling rates, such as minutes, possibly even hour or up to 8 hours in relation to the actual chemical and metallurgical processes. Large variances in time shifts between these different attributes lead to the temperature of the actual tapped iron, namely the Hot Metal Temperature (HMT), being contextually related to a range of process attributes variously distributed backwards in time.

Various researchers have previously attempted to utilise Artificial Neural Networks (ANNs), because of their computational efficiency and little to no requirement of prior domain knowledge (Rumelhart et al. 1988), for modelling the complex inter-variable relationships, such as predicting the silicon content of molten iron (Banks 1999, Sarma 2000). Bhattacharjee et al. (1999) proposed traditional feed-forward neural networks to predict a number of quality parameters of the molten iron, including a categorised HMT. This work indicated that the various multi-layer perception networks were able to predict daily HMT trends utilizing only 15 inputs from the blast furnace. The extreme learning machine (ELM) algorithm was also used for generating dynamic modelling for a predictive model that targeted the silicon content level in the molten iron (Zhou et al. 2015).

In addition to ANN algorithms, other techniques have also been evaluated in the iron-making industry for various purposes. Kommenda et al. (2011) used an unguided symbolic regression approach for variable selection and knowledge extraction from BF dataset. In this work, Genetic Programming (GP) (Koza 1992) was executed multiple times for reducing the stochastic effects and for identifying important variables through a variable interaction network. In a recent work proposed by Harvey and Gheribi (2014), a direct search algorithm called MADS was combined with the classical thermodynamics modelling of for optimising the process parameters of a BF. Promising outcomes were produced in the BF simulation, and the model indicated improved precision in subsequent quantitative analytic task, such as the degradation of the BF refractory materials and liquid metal quality control. A Kalman filter and minimum description length (MDL) algorithm was also employed by Waller et al. in a series of works (2000, 2002), in order to improve the linear ARMA and FIR models for predicting the silicon content in the hot metal produced by a BF.

In considering general TSDM work, much of the assessed research focuses on producing appropriate time series (TS) representations in order to retain the order of the data. However, these representations, considered as TS features, are further analysed by a limited number of simpler methods. Some TSDM approaches rely on the indexability of their features (Shieh and Keogh 2008, Agrawal et al. 1993), whereas others focused on similarity based features (Möller-Levet et al. 2003, Morrill 1998).

Beyond the level of representations, orders among nominal features can be represented by Allen's Interval Algebra (Allen 1983), in which 13 basic relations between two intervals are defined, or various extensions of Allen's definition (Freksa 1992, Roddick & Mooney 2005). However, with numerous potential ordering types between intervals, the problem becomes significantly more complex in a real time series dataset with tens or hundreds of representative features.

Inference of the cause-effect relationships is commonly based on the Granger Causality (Granger 1969, 1980), which evokes two fundamental principles: (1) the effect does not precede its cause in time; (2) the casual series contains unique information about the series being caused that is not available otherwise. Although Granger Causality was initially proposed for solving economics problems, it has been widely utilized for mining time series data. Qiu et al. (2012) used Granger graphical models to compute the correlation anomaly of each variable in various industrial time series. Mohammad and Nishida (2012) utilised the Granger causality for discovering the casual structure of interesting recurring events in multi-dimensional time series data. The same method was also used for determining the size of sliding windows and feature selection in multivariate time series with lagged values (Sun et al. 2015).

Episode mining (Mannila et al. 1995) is a specialised approach for analysing temporal event data with numerous event combinations that occur within a given window. This technique can be extended to general time series data if nominal TS features are defined as events. The concept of an episode emphasises together with the importance of combinations of various events, that this conforms to the manner in which humans often perceive temporal events (Batyrshin & Sheremetov 2008). Despite this extensive research on episode mining considers the discovery of frequent serial episodes as the fundamental problem (Mannila et al. 1997, Laxman et al. 2007), as these frequent episodes are believed to have a higher importance than infrequent episodes. However, in the real world, this is not always true. For example, a combination that repeatedly occurs may not be interesting to the observer.

A similar situation arises with decision tree modelling. Typically, every node in a tree represents a conditional test out of many others which divides the data into segments with a purer mixture of classes at each descending level. Therefore, the meaningfulness of a condition node is measured by the information gain obtained in dividing the dataset. If all event combinations in an event sequence are labelled, then the meaningfulness of these can be measured in a similar way. In this paper, a new sequential data classification approach is presented where the meaningful event combinations are the key factors. The meaningfulness of a combination is established by an XoN (X-of-N) decision tree (Zheng 2000) rather than by the statistical measure, therefore this approach is based on information theory. Multiple event groups are constructed and selected in order to form compact XoN features that are used in the tree model.

An XoN representation contains special features that cover multiple possible combinations of given conditions. This is formed by two parts, X and N, and each is a non-empty unsorted set. The N component consists of several traditional conditions (TS events in this case), and the X component is a list of non-negative integers, denoting the exact numbers of conditions in N to be satisfied.

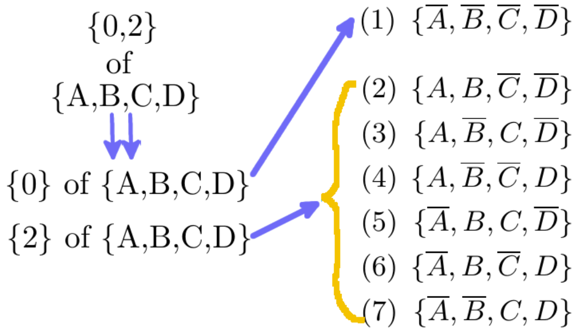

{0,2}
of
{A,B,C,D}
{0} of {A,B,C,D}
{2} of {A,B,C,D}
(1) $\{\overline{A},\overline{B},\overline{C},\overline{D}\}$
(2) $\{A,B,\overline{C},\overline{D}\}$
(3) $\{A,\overline{B},C,\overline{D}\}$
(4) $\{A,\overline{B},\overline{C},D\}$
(5) $\{\overline{A},B,C,\overline{D}\}$
(6) $\{\overline{A},B,\overline{C},D\}$
(7) $\{\overline{A},\overline{B},C,D\}$

*Figure 3: A full expansion of an XoN feature.*

A typical XoN feature can be decomposed into a number of combinations of various unordered events, as seen in Figure 3. The simple XoN representation covers seven different situations, where the A, B, C and D denote conditions that may occur concurrently in the dataset. Although the XoN algorithm was not developed for time series analysis, we understand there are similarities between the decomposed event combinations and episodes. As an integrated feature, an XoN node practically gathers a series of episodes and combines their classifying capacities in order to obtain a better performance.

## 3 Methodology

The method is designed to process sequential data with either numeric or nominal values, and is based on an idea that any real-world event can be observed and identified as a combination of other observable "events" over various temporal attributes. The word "event" here is defined as any distinguishable feature derived locally from a sequential data, such that, it may refer to temporal shapes, different value zones, transformations of the real data, symbol segments or any other features that reflect the characteristics of any temporal segments.

Whilst not all contextual features are tied to targeted real-world events, some feature combinations across various attributes however, may be exclusively associated with such an event. This is similar to the manner in which a doctor may diagnose patients: one symptom alone may not be sufficient evidence of a certain disease, however a set of relevant symptoms could align the diagnosis to a specific disease with considerable confidence. In this paper, an EGBC framework, the "events" hold an equivalent meaning to the "symptoms" in a medical diagnosing system.

The basic assumption can be described as follows: If a (detectable) process occurs at some time point *T*, its causal factors or resultant responses can be represented as time series features and in turn transformed into nominal events that occur around the time point *T*. The novelty of this work is that all casual and consequent events within a given time frame are indiscriminately analysed as unsorted combinations, and the type of the target process can be classified based on these combinations, so that the potential localised temporal misalignments are eliminated within the given time frame. Therefore, in situations where specific classes are clearly labelled on each stage of such sequences, the goal is to build a model that classifies these segments based on certain important event combinations they contain. Most time series data are not originally presented in the form of nominal events, thus an appropriate transformation is required to represent the time series in a nominal event form. Event groups are extracted with a sliding window over the sequence, and meaningful combinations of these are automatically constructed and selected by an adapted XoN algorithm.

The flow chart in Figure 4 briefly explains the overall procedure with a simplified example. Assuming there are three different working stages in a process[1], and the task is then set to

[1] For instance, the HMT may be at higher, lower or normal states from the expected temperature level during different periods, and these periods can be labelled as three classes (High, Low, Normal) which are associated to different BF working stages respectively.

identify these stages (classes) from two associated time series TS-A and TS-B. The numeric time series are then transformed into two event sequences respectively in which synchronisation is not compulsory, named Seq-A and Seq-B, and events are labelled as An and Bn accordingly (n being an event numbers). Event groups are then formed across both sequences by selecting all events that fall within a sliding window, and labelled by the classes as the training dataset. The XoN decision tree model is then trained in order to find the optimised event combinations for classifying the various process stages. During this procedure, an understanding of the expected event combinations and their correct labelling are essential for verifying whether this approach is effectively finding useful combinations and classifying the data accurately.

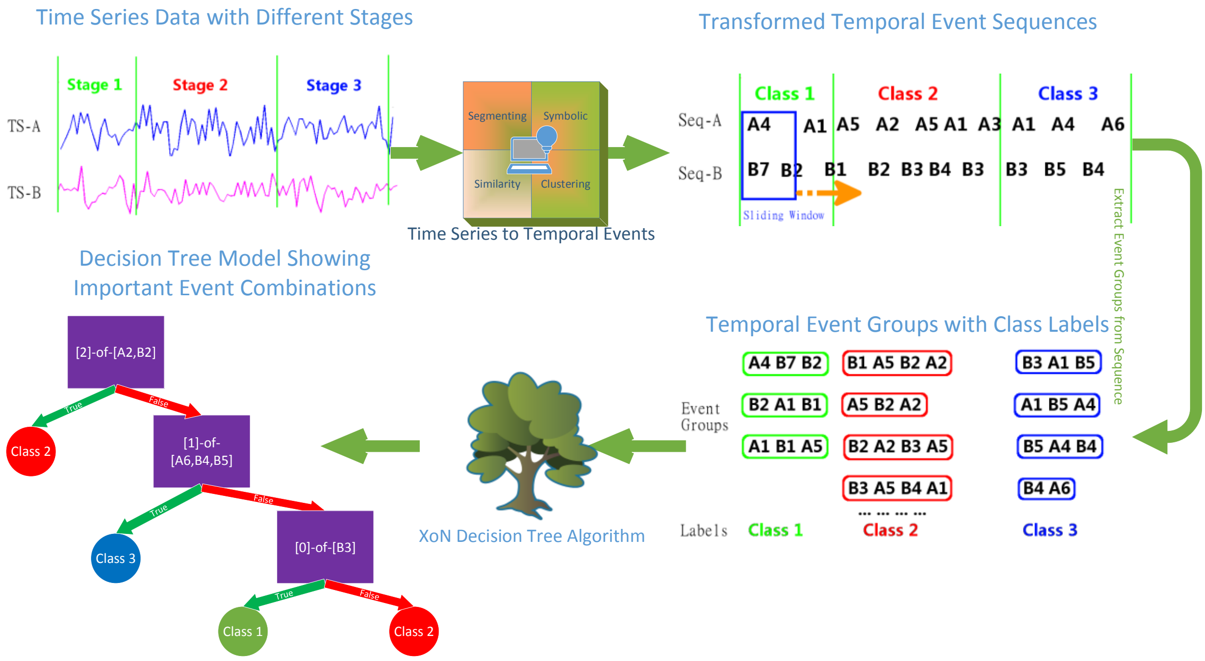


*Figure 4: The Procedure of Event Group Based TS Classification Approach.*

In order to evaluate the feasibility of our approach, we employed our event-group based approach for two different sequential data classification tasks in the following experiment section, described as follows:

1. **Language identification**. The goal of this task is to recognise the languages from three articles written in different languages. Our approach is used to identify what language a word is written in without any dictionary. This task demonstrates how our approach performs on real event based sequences (non-time series) rather than artificially generated data. The outcomes are compared with the text mining algorithm – "TextCat" (Hornik et al. 2013).

2. **Honeybee Dance Behaviour Recognition**. In order to identify the dancing behaviours of a honey bee, genuine honeybee motion data are studied by employing our event group based approach. The sequential real-valued bee motion data are transformed and clustered into events before being analysed by the XoN decision trees. This task extends the online classification work to a real-world scenario, involving multi-dimensional time series data.

3. **Hot Metal Temperature Prediction**. The aim of this task is to model the blast furnace process based on the abundant historical operational data in order to predict if the future HMT falls outside a normal range. Multiple TS attributes are included in the dataset with various connotations and sampling rates, among which the temporal misalignment issue exist. The events in the BF data are based on the basic shapes, and ensemble of XoN models are generated and updated regularly.

# 4 Experiment And Results

In this section, the method proposed in Section 3 will initially be applied on two sequential classification tasks, and subsequently for modelling a more significantly complex industrial problem. A language identification (LID task) is conducted in order to model and distinguish three different languages (English, Italian and Dutch) without the benefit of any prior linguistic or dictionary based knowledge. With naturally occurring events and groups within the text, i.e. letters and words, the aim of this task is to verify the feasibility of our event group based methodology. In the second, the honey bee dancing task, we evaluate the method on real-valued time series. In both tasks, online classification along continuous sequences is accomplished, i.e. different classes in different states, and in a progressive manner in terms of the complexity. Finally, the evaluated full EGBC framework is utilised for modelling the iron-making dataset in order to predict abnormal states of the blast furnace.

## 4.1 Text Mining – Language Identification

The automatic identification of a language from text is an important and well-studied problem, which was considered as a solved problem (McNamee 2005). LID is a typical multi-class classification task, however, in the past, decision tree techniques were not considered in this domain. The general LID methods are based on either statistical language modelling or the frequency of common word usage. Both the statistical and frequency of common words usage methods work better on sentences that consist of more than 15 words (McNamee 2005). The accuracy of various traditional LID methods approaches some 98% if the decision is based on sentences or short paragraphs (Takcı & Soğukpınar 2004).

For the general LID task, identifying the language of single words is not practically necessary, because it is not common to mix different languages within individual sentences and paragraphs. However, in this work, in order to evaluate the proposed event group based classification approach, all articles are analysed and classified word by word.

### 4.1.1 Dataset

Six public domain e-books in the three specific languages are obtained from the Project Gutenberg website (*Project Gutenberg*. n.d.), two in each language[2]. In order to ensure that only letter combinations are used as a valid basis for the classification process, all symbols and non-English characters are removed. Words with three or fewer letters are ignored and the texts are pre-processed with the Porter Stemming algorithm (Porter 1980) to reduce the word form variances for better information retrieval. The sizes of the training and testing datasets are 118,192 words (Eng: 31,387; Ita: 50,282; Dut: 36,523) and 76,143 words (Eng: 25,644; Ita: 34,068; Dut: 16,431) respectively.

Because words are the basic meaningful elements in languages, the event groups are naturally defined as single words. As the spelling is sensitive to the order of characters, in order to partially retain ordering information, every pair of adjacent characters in a word are defined as events rather than individual letters. Preliminary experiments indicate that the XoN models using 2-letter events have a noticeable improvement (5%) on classification accuracy compared to models using 1-letter events, if the same modelling parameters are used.

### 4.1.2 Experiment Setting and Results

Following the pseudocode in Zheng's paper, the XoN algorithm was implemented using Python 2.7. In order to check the stability of the method, a number of XoN models were generated based on a random 60% of the training text, and all performed similarly in terms of the amount

[2] Training Texts: English: http://www.gutenberg.org/files/1999/1999.txt
Italian: http://www.gutenberg.org/cache/epub/1012/pg1012.txt
Dutch: http://www.gutenberg.org/files/18066/18066-8.txt
Testing Texts: English: http://www.gutenberg.org/ebooks/43230
Italian: http://www.gutenberg.org/ebooks/43226
Dutch: http://www.gutenberg.org/ebooks/11500

of pruning used and their accuracy. In the following experiments, the model with the highest classification accuracy and moderate pruning was chosen for further analysis.

In this experiment, the XoN tree model is not forced to make a decision on every event group (word), as some words may exist in different languages, and these cross-class groups are not classifiable. For the aim of either, classifying languages or obtaining linguistic knowledge, the classification of every word is not necessary. In fact, ignoring some common words is seen to help the model to focus on the more important linguistic characteristics.

In order to avoid vague decisions, in this experiment, a minimum confidence parameter is used to control the output behaviour of the XoN model. When the confidence from a decision is lower than a given threshold "minConf", the classifier output is altered to "unknown" rather than the most likely class. Introducing this "unknown" class and a minimum confidence reduces the number of overall mis-classifications and increases the overall classification accuracy. Another tree pruning parameter is the general concept of "minCase", minimum number of instances in a node before stopping further splitting. Figure 5 illustrates how the accuracies and unknown ratios (the number of unknown cases divided by the size of the test dataset) are affected when the model is being pruned with various minCase and minConf.

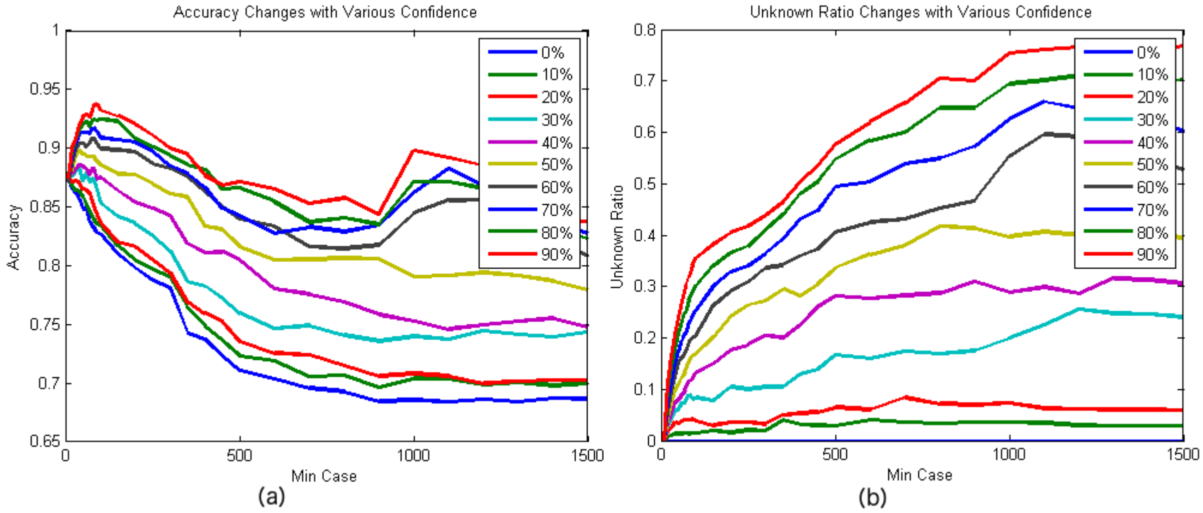


*Figure 5: Classification Performance on Models with Different minCase and minConf. a) Accuracy; b) Unknown Ratio.*

In Figure 5, as the minimum confidence varies from 0% to 90%, the classification accuracy on the testing text increases. However, a high confidence requirement also renders the model as ineffective when the tree is heavily pruned, and the unknown ratio becomes too high. It is reasonable to assume that an optimal model could be expected to maintain a relatively high accuracy whilst keeping the unknown ratio below an acceptable level.

The yellow line (minConf = 50%) in Figure 5(a) reaches its peak (accuracy = 89.6%) when the value of minCase is 40, and the unknown ratio for the same model is 9.8% when classifying the testing dataset. Considering the highest possible accuracy model has 93.79% accuracy and a 34.09% unknown ratio, the yellow line model is considered to be an optimised tree model for the task. Note that the training and testing texts are sourced from different e-books with different authors, styles and expressions. It is also expected that this approach would also improve if the training and testing data were sourced from the same article or author context.

### 4.1.3 Performance Comparison

Text-Cat (van Noord n.d.), a LID implementation of the N-gram-based text categorization (Cavnar et al. 1994), is applied on the testing dataset as a comparison. The Text-Cat is limited to the selected three languages in this experiment although it supports up to 69 different languages. Text-Cat may make multiple decisions on a word, and these are post-processed to either "unknown" or a false decision. For example, if an English word is identified as "English or Italian", then the result is converted to an "unknown"; and if an English word is identified as "Italian or Dutch", because both are incorrect, the first incorrect decision will be kept.

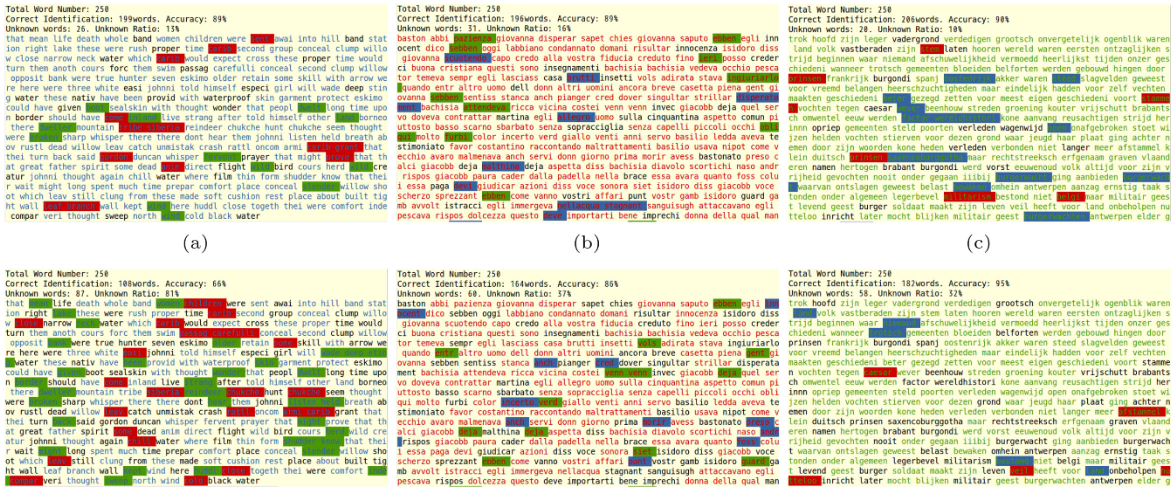
Table 1: Identification on Testing Texts with XoN and Text-Cat Models. Based on different pruning options, the same fully grown XoN tree model has various sizes (node number), accuracies and unknown ratios, listed as Model 1, 2 and 3 respectively.

| Model | *XoN Model 1*<br>minCase= 40 minConf= 50% | *XoN Model 2*<br>minCase= 40 minConf= 98% | *XoN Model 3*<br>minCase= 460 minConf= 50% | *Cat Model*<br>Original |
|---|---|---|---|---|
| Size | 1222 Nodes | 1342 Nodes | 292 Nodes | 1200 Terms |
| Accuracy | 89.6% | 91.9% | 83% | 81.7% |
| Unknown | 0 9.9% 1 | 0 25% 1 | 0 32.6% 1 | 0 32.4% 1 |
| Confusion Matrix (%) | *Pred!* *Eng* *Ita* *Dut*<br>B *Eng* 88.9 7.3 3.8 C<br>@ *Ita* 6.5 90.9 2.6 A<br>*Dut* 6.6 5.0 88.4 | *Pred!* *Eng* *Ita* *Dut*<br>B *Eng* 90.1 6.5 3.4 C<br>@ *Ita* 5.0 93.5 1.5 A<br>*Dut* 5.5 3.8 90.7 | *Pred!* *Eng* *Ita* *Dut*<br>B *Eng* 70.4 23.4 6.2 C<br>@ *Ita* 6.6 89.6 3.8 A<br>*Dut* 9.3 5.9 84.8 | *Pred!* *Eng* *Ita* *Dut*<br>B *Eng* 60.1 19.4 20.5 C<br>@ *Ita* 6.9 89.1 4.0 A<br>*Dut* 6.1 1.8 92.1 |

*Table 1: Identification on Testing Texts with XoN and Text-Cat Models. Based on different pruning options, the same fully grown XoN tree model has various sizes (node number), accuracies and unknown ratios, listed as Model 1, 2 and 3 respectively.*

The Text-Cat models used in this work are publicly available (van Noord n.d.), and each language model file contains 400 high frequency terms. A comparison between Text-Cat and the XoN tree models (with three different pruning options) can be viewed in Table 1. The first XoN decision tree model is lightly pruned, containing some 1222 nodes and achieves 89.6% word-by-word classification accuracy on the testing texts (9.8% unknown words of the testing texts). As a comparison, the Text-Cat has 81.7% in accuracy on the same testing text (32.4% unknown words of the testing texts). Confusion matrices expressed in percentages, are also included in Table 1 for easy comparison between all models.

One advantage of the XoN decision tree model is that it can be easily pruned to suit different requirements. For instance, if the accuracy is of higher priority than the unknown ratio, the model can be pruned as Model 2 in Table 1, which provides a similar unknown ratio to what Text-Cat provides, but with a significantly higher accuracy at 91.9%. However, if a smaller, less complex form is preferred, the XoN model can also be further pruned for less nodes. With a minCase=460 and minConf=50%, such as Model 3 which contains only 292 nodes yet also produces a similar unknown ratio (32.6%) compared to Text-Cat. This model still provides a marginally higher identification accuracy at 83% on the test dataset.



```
Total Word Number: 250
Correct Identification: 199words. Accuracy: 89%
Unknown words: 26. Unknown Ratio: 13%
that mean life death whole band women children were sent awai into hill band stat
ion right lake these were rush proper time carib second group conceal clump willo
w close narrow neck water which carib would expect cross these proper time would
turn them anoth cours forc them swim passag carefulli conceal second clump willow
 opposit bank were true hunter seven eskimo older retain some skill with arrow we
re here were three white easi johnni told himself especi girl will wade deep stin
g water these nativ have been provid with waterproof skin garment protect eskimo
could have given boot sealskin with thought wonder that peopl dwelt long time upo
n border should have come inland live strang after told himself other land borneo
 there dweller mountain tribe siberia reindeer chukche hunt chukche seem thought
were broken sharp whisper there there dont hear them johnni listen held breath ab
ov rustl dead willow leav catch unmistak crash rattl oncom armi carib grant that
thei turn back said gordon duncan whisper fervent prayer that might prove that th
at great father spirit some dead anim direct flight wild bird cours herd wild cre
atur johnni thought again chill water where film thin form shudder know that thei
r wait might long spent much time prepar comfort place conceal slender willow sho
ot which leav still clung from these made soft cushion rest place about built tig
ht wall leaf branch wall kept wind here huddl close togeth thei were comfort inde
 compar veri thought sweep north wind cold black water
```

(a)

```
Total Word Number: 250
Correct Identification: 196words. Accuracy: 89%
Unknown words: 31. Unknown Ratio: 16%
baston abbi pazienza giovanna disperar sapet chies giovanna saputo ebben egli inn
ocent dico sebben oggi labbiano condannato domani risultar innocenza isidoro diss
 giovanna scuotendo capo credo alla vostra fiducia creduto fino ieri posso creder
ci buona cristiana questi sono insegnamenti bachisia bachisia vedeva occhio pesca
tor temeva sempr egli lasciass casa brutti insetti vols adirata stava ingiuriarlo
 quando entr altro uomo dell donn altri uomini ancora breve casetta piena gent gi
ovanna sebben sentiss stanca anch pianger cred dover singultar strillar disperata
ment bachisia attendeva ricca vicina costei venn venn invec giacobb deja quel ser
vo doveva contrattar martina egli allegro uomo sulla cinquantina aspetto comun pi
uttosto basso scarno sbarbato senza sopracciglia senza capelli piccoli occhi obli
qui molto furbi color incerto verd giallo venti anni servo basilio ledda aveva te
stimoniato favor costantino raccontando maltrattamenti basilio usava nipot come v
ecchio avaro malmenava anch servi donn giorno prima morir avess bastonato preso c
alci giacobb deja malthina deja aspetta diss bachisia diavolo scortichi naso andr
 rispos giacobb paura cader dalla padella nella brace essa avara quanto foss colu
i essa paga devi giudicar azioni diss voce sonora siet isidoro diss giacobb voce
scherzo sprezzant ebben come vanno vostri affari punt vostr gamb isidoro guard ga
mb avvolt istracci egli immergeva nellacqua stagnant sanguisugh attaccavano egli
pescava rispos dolcezza questo deve importarti bene imprechi donna della qual man
```

(b)

```
Total Word Number: 250
Correct Identification: 206words. Accuracy: 90%
Unknown words: 20. Unknown Ratio: 10%
trok hoofd zijn leger vadergrond verdedigen grootsch onvergetelijk ogenblik waren
 land volk vastberaden zijn stem laten hooren wereld waren eersten ontzaglijken s
trijd beginnen waar niemand afschuwelijkheid vermoedd heerlijkst tijden onzer ges
chiedeni wanneer trotsch gemeenten bloeiden belforten werden gebouwd hingen door
prinsen frankrijk burgondi spanj oostenrijk akker waren steed slagvelden geweest
voor vreemd belangen heerschzuchtigheden maar eindelijk hadden voor zelf vechten
maakten geschiedeni beter gezegd zetten voor meest eigen geschiedeni voort stamme
n vochten tegen caesar wever beenhouw streden groening kouter vrijschutt brabants
ch omwentel eeuw werden factor wereldhistori kone aanvang reusachtigen strijd her
innn opriep gemeenten steld poorten verleden wagenwijd open onafgebroken stoet wi
jzen helden vochten stierven voor dezen grond waar jeugd haar plaat ging achter n
emen door zijn woorden kone heden verleden verbonden niet langer meer afstammel k
lein duitsch prinsen saxencoburggotha maar rechtstreeksch erfgenaam graven vlaand
eren namen hertogen brabant burgondi werd vorst eeuwenoud volk altijd voor zijn v
rijheid gevochten nooit onder gegaan iiibij burgerwacht ging aanbieden burgerwach
t waarvan ontslagen geweest belast bewaken omhein antwerpen aanzag ernstig taak s
tonden onder algemeen legerbevel militarism bestond niet belgi maar militair gees
t levend geest burger soldaat maakt zijn leven veil heeft voor land onbeholpen nu
tteloo inricht later mocht blijken militair geest burgerwachten antwerpen elder g
```

(c)

```
Total Word Number: 250
Correct Identification: 108words. Accuracy: 66%
Unknown words: 87. Unknown Ratio: 81%
that mean life death whole band women children were sent awai into hill band stat
ion right lake these were rush proper time carib second group conceal clump willo
w close narrow neck water which carib would expect cross these proper time would
turn them anoth cours forc them swim passag carefulli conceal second clump willow
 opposit bank were true hunter seven eskimo older retain some skill with arrow we
re here were three white easi johnni told himself especi girl will wade deep stin
g water these nativ have been provid with waterproof skin garment protect eskimo
could have given boot sealskin with thought wonder that peopl dwelt long time upo
n border should have come inland live strang after told himself other land borneo
 there dweller mountain tribe siberia reindeer chukche hunt chukche seem thought
were broken sharp whisper there there dont hear them johnni listen held breath ab
ov rustl dead willow leav catch unmistak crash rattl oncom armi carib grant that
thei turn back said gordon duncan whisper fervent prayer that might prove that th
at great father spirit some dead anim direct flight wild bird cours herd wild cre
atur johnni thought again chill water where film thin form shudder know that thei
r wait might long spent much time prepar comfort place conceal slender willow sho
ot which leav still clung from these made soft cushion rest place about built tig
ht wall leaf branch wall kept wind here huddl close togeth thei were comfort inde
 compar veri thought sweep north wind cold black water
```

(d)

```
Total Word Number: 250
Correct Identification: 164words. Accuracy: 86%
Unknown words: 60. Unknown Ratio: 37%
baston abbi pazienza giovanna disperar sapet chies giovanna saputo ebben egli inn
ocent dico sebben oggi labbiano condannato domani risultar innocenza isidoro diss
 giovanna scuotendo capo credo alla vostra fiducia creduto fino ieri posso creder
ci buona cristiana questi sono insegnamenti bachisia bachisia vedeva occhio pesca
tor temeva sempr egli lasciass casa brutti insetti vols adirata stava ingiuriarlo
 quando entr altro uomo dell donn altri uomini ancora breve casetta piena gent gi
ovanna sebben sentiss stanca anch pianger cred dover singultar strillar disperata
ment bachisia attendeva ricca vicina costei venn venn invec giacobb deja quel ser
vo doveva contrattar martina egli allegro uomo sulla cinquantina aspetto comun pi
uttosto basso scarno sbarbato senza sopracciglia senza capelli piccoli occhi obli
qui molto furbi color incerto verd giallo venti anni servo basilio ledda aveva te
stimoniato favor costantino raccontando maltrattamenti basilio usava nipot come v
ecchio avaro malmenava anch servi donn giorno prima morir avess bastonato preso c
alci giacobb deja malthina deja aspetta diss bachisia diavolo scortichi naso andr
 rispos giacobb paura cader dalla padella nella brace essa avara quanto foss colu
i essa paga devi giudicar azioni diss voce sonora siet isidoro diss giacobb voce
scherzo sprezzant ebben come vanno vostri affari punt vostr gamb isidoro guard ga
mb avvolt istracci egli immergeva nellacqua stagnant sanguisugh attaccavano egli
pescava rispos dolcezza questo deve importarti bene imprechi donna della qual man
```

(e)

```
Total Word Number: 250
Correct Identification: 182words. Accuracy: 95%
Unknown words: 58. Unknown Ratio: 32%
trok hoofd zijn leger vadergrond verdedigen grootsch onvergetelijk ogenblik waren
 land volk vastberaden zijn stem laten hooren wereld waren eersten ontzaglijken s
trijd beginnen waar niemand afschuwelijkheid vermoedd heerlijkst tijden onzer ges
chiedeni wanneer trotsch gemeenten bloeiden belforten werden gebouwd hingen door
prinsen frankrijk burgondi spanj oostenrijk akker waren steed slagvelden geweest
voor vreemd belangen heerschzuchtigheden maar eindelijk hadden voor zelf vechten
maakten geschiedeni beter gezegd zetten voor meest eigen geschiedeni voort stamme
n vochten tegen caesar wever beenhouw streden groening kouter vrijschutt brabants
ch omwentel eeuw werden factor wereldhistori kone aanvang reusachtigen strijd her
innn opriep gemeenten steld poorten verleden wagenwijd open onafgebroken stoet wi
jzen helden vochten stierven voor dezen grond waar jeugd haar plaat ging achter n
emen door zijn woorden kone heden verleden verbonden niet langer meer afstammel k
lein duitsch prinsen saxencoburggotha maar rechtstreeksch erfgenaam graven vlaand
eren namen hertogen brabant burgondi werd vorst eeuwenoud volk altijd voor zijn v
rijheid gevochten nooit onder gegaan iiibij burgerwacht ging aanbieden burgerwach
t waarvan ontslagen geweest belast bewaken omhein antwerpen aanzag ernstig taak s
tonden onder algemeen legerbevel militarism bestond niet belgi maar militair gees
t levend geest burger soldaat maakt zijn leven veil heeft voor land onbeholpen nu
tteloo inricht later mocht blijken militair geest burgerwachten antwerpen elder g
```

(f)

*Figure 6: Word-by-Word Language Identification by XoN and Text-Cat Models. Font colours indicate True Class, Background colours indicate Identified Class. Meaning of Colours: Blue-English, Red-Italian, Green-Dutch, Black-Unknown. (a,b,c) Results from XoN Model; (d,e,f) Results from Text- Cat Model.*

The actual XoN tree models in Table 1 are too large to be included in this paper. Figure 6 illustrates a number of the classification results where three sections of texts and their classification results are visually presented. In Figure 6, the English, Italian and Dutch are printed in blue, red and green fonts. If a word is mis-classified, the background colour of that word changes to the colour of the wrong decision accordingly. Further, when an unknown identification is made, the word is presented in black.

## 4.2 Honeybee Dancing Behaviour Classification

In this task, classification on genuine, real-valued time series data is involved. The honey bee dancing dataset (Oh et al. 2008) includes six videos of honeybee dancing, where the trajectory of a signalling bee is automatically tracked and converted into quantitative sequential motion data, including sequences of position (X and Y) and the head angle of the bee. This dataset contains ground truth labels of all data records according to which behaviour of three possible behavioural patterns it contains: waggle, right-turn or left-turn. A number of researchers have previously attempted to classify the type of dancing behaviour based on these motion features, using extended HMM, segmentation (Fox et al. 2008) and clustering methods (Zhou et al. 2013). In this section, the same classification task is performed to show how our event based decision tree method performs on this same dataset.

### 4.2.1 Data Pre-processing

In order to transform the real-valued motion data into nominal events, we define patterns of movements and consequential movements at any given time point *T* as the event for describing the dancing behaviours. For example, it is safe to assume when the behaviour is left-turn, the bee's position moves towards the left relative to the previous trajectory. For the waggle behaviour, the bee may move both left and right in an alternating manner. Therefore, the basic movement patterns for this task are closely associated with the directions of movements. The sequences of raw positions and head angles are transformed into the following six features:

- *X-Off*: The X offset of current position relative to the previous frame.
- *Y-Off*: The Y offset of current position relative to the previous frame.
- *CosAng*: Cosine of the head angle.
- *SinAng*: Sine of the head angle.
- *AngDiff*: The difference between head angle and previous trajectory.
- *Dist*: The distance the bee travels since last frame.

In the above list, the *CosAng* and *SinAng* features can be considered as an estimation of the *X-Off* and *Y-Off* for the following moment, if the bee maintains its current heading direction and speed. The *AngDiff* indicates the difference between the current heading and that of the previous movement. From every frame, a vector of all six features is viewed as a unique motion status of the dancing bee, and every status contains information derived from both the current and previous frame.

If every such motion status is treated as a nominal event, there will be a massive number of events. The curse of dimensionality will affect our decision tree algorithm that would seek to find the optimised combinations. The feature vectors are therefore pre-clustered using a Minimum Message Length (MML) algorithm (Wallace & Boulton 1968). Subsequently, each cluster is treated as an event representing a series of similar motion states. The MML clustering also reduces the six-dimensional feature sequences into a single dimensional event sequence with a limited number of events.

### 4.2.2 Parameter details

Unlike the textual LID event sequences, the honeybee dancing events have uniform time intervals, therefore a fixed number of events are selected with a fixed-width sliding window. The behavioural label of an event group is determined by the majority of the ground truth within that window. In order to obtain an optimised model reflecting the inherent truth within the honey bee dancing motion data, a series of training processes were executed with various parameters, such as different initialisations, training datasets, numbers of MML clusters, sliding window sizes and the minimum case numbers required before splitting a node. The parameters used in this experiment are listed below:

- Number of MML clusters: This parameter was pre-set to be 8, 10 or 12 clusters.
- Sliding window size (WinSize): 5,7,9,11. These are also the sizes of event groups, i.e. the number of continuous video frames used for classifying a behaviour. Only an odd number of events are used for simple majority labelling.
- XoN seed: 100, 200, 300, 400. This seed controls all random functions in the XoN training procedure, such as initialisation, dataset division etc.
- Training ratio: 70% of the sequential data are used for training process and the rest are for testing purpose. Because the continuity is important in the time series data, the testing data is always sourced from a continuous section of the sequence.
- Maximum X (Max-X): Varies from the number four (4) up to the number (WinSize-1). Because the event group size is WinSize, any X greater than WinSize is meaningless, therefore the maximum integer in the X part is limited by the number (WinSize -1).
- Minimum case number for splitting a node (MinCase): 2,5,8,10,15,20,40,70. These are traditional parameter values used in many decision trees, and indicates the minimum size of data before a node can be further split, to extend the model.

### 4.2.3 Model Selection

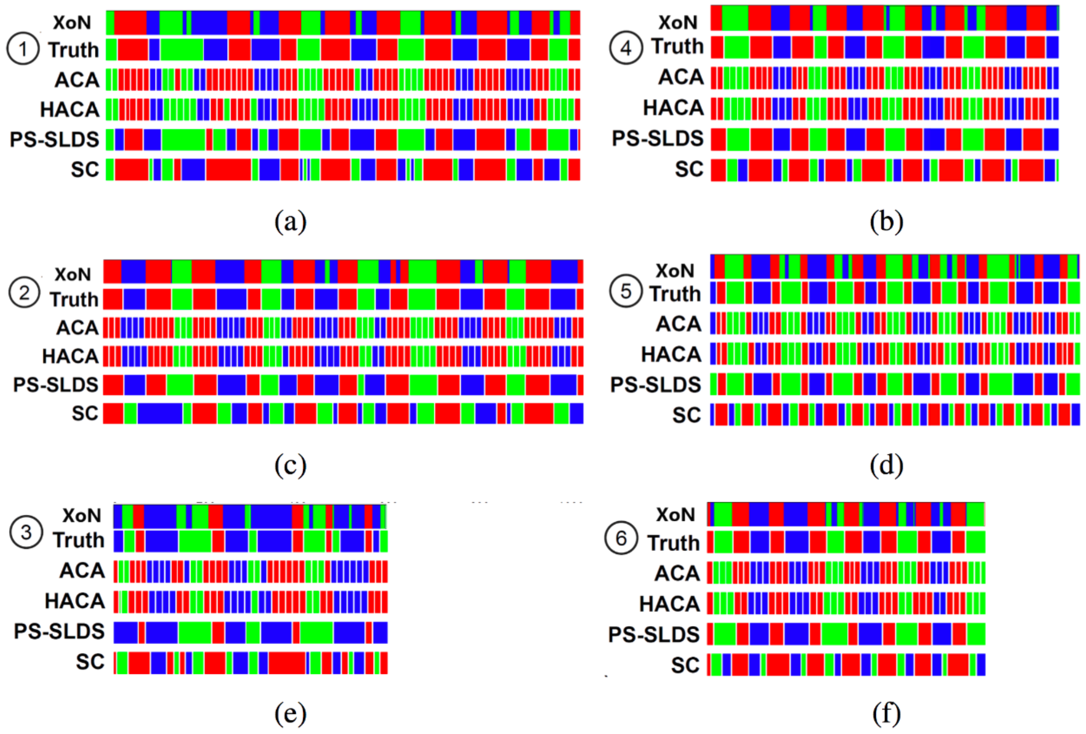
The training process of the XoN algorithm exhaustively uses all parameters listed in previous section, and the models are evaluated on all three datasets (training, testing and whole) for error rates.

| id | cNo | Seed | winSize | Max-X | MinCas | slideAcc | ErrDiff |
|---|---|---|---|---|---|---|---|
| 1 | 12 | 400 | 11 | 5 | 15 | 85.3% | 0.4% |
| 2 | 8 | 300 | 11 | 8 | 5 | 93.0% | 0.7% |
| 3 | 8 | 100 | 9 | 7 | 20 | 87.04% | 9.5% |
| 4 | 14 | 400 | 11 | 5 | 10 | 90.62% | 7.7% |
| 5 | 16 | 200 | 11 | 8 | 15 | 88.92% | 11.5% |
| 6 | 14 | 200 | 9 | 5 | 15 | 88.17% | 12.1% |

*Table 2: Selected models and sliding prediction accuracy.*

The classification error rates are calculated through two techniques: a raw prediction error rate and a sliding prediction error rate. The raw error rate is the case by case error rate showing the accuracy of event group classification. However, because each frame of data is contained in multiple event groups, the classification on a single frame should also be further determined using the majority of classifications it receives as the sliding window passes. The sliding error rate in general is about 5-10% better than the raw prediction error rate, thus all error rates or accuracies in the rest of this section are based on the sliding method. In most cases, a fully grown decision tree model would have higher classification accuracy on the training data than on the testing data, this is of course a symptom of over training as the model performs worse on previously unseen data. Pruning with larger values of MinCase normally reduces this performance difference, however it also lowers the overall accuracy. In order to select an optimised tree model from all the trained models, the models are ascendingly sorted based on a score = AllErr + abs(TrainErr − TestErr). A good model is expected to have a low score indicating that both the overall error rate and the performance difference between testing and training datasets are low.

### 4.2.4 Experimental Results



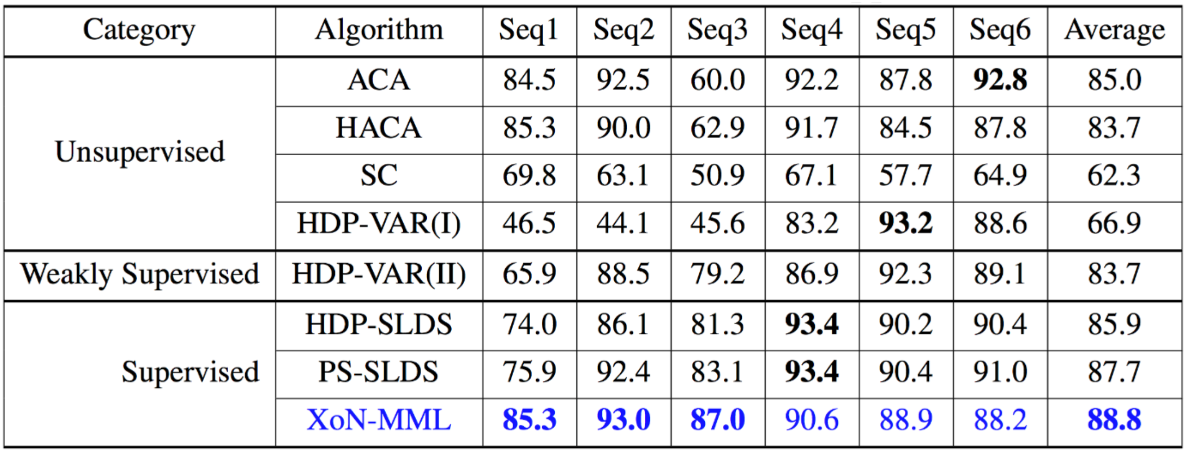
(a) (b) (c) (d) (e) (f)

*Figure 7: Comparison with classification result from other researches; where the motions are waggle (green), right-turn (red), left-turn (blue). (a) – (f) show outcomes for 6 individual honeybee motion sequences.*

The exhaustive training process selects the following model parameters for classifying each honeybee dancing data as shown in Table 2. Table 2 illustrates that the best fully grown XoN tree models discovered during the exhaustive training process. Due to the limited number of models that were generated, these are not reflective of the best performance our approach could achieve. Even though, the performances of these models are comparable and even exceed several of other techniques which were employed for the same task. Zhou et al. (Zhou et al. 2013) summarised a table to compare some state of the art techniques for classifying the honeybee dancing dataset, and the accuracies are further compared with our method in Figure 7.

### 4.2.5 Discussion

The accuracy comparison in Table 3 indicates that selected event group based decision tree models have comparable classification performance to other existing techniques. Our method outperforms all other techniques in three of the six sequences, and it also has the highest average accuracy in all of the compared methods.

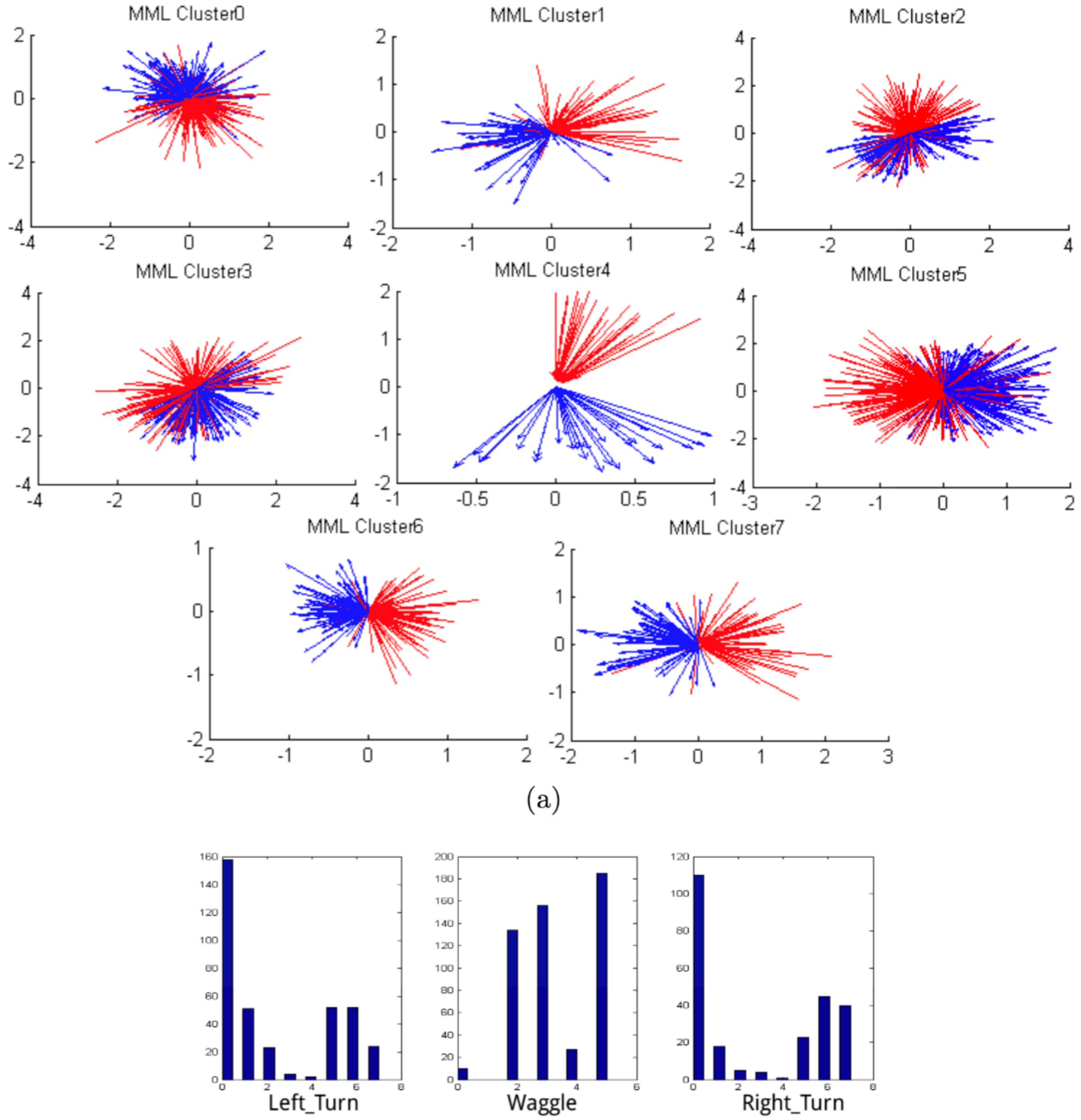


| Category | Algorithm | Seq1 | Seq2 | Seq3 | Seq4 | Seq5 | Seq6 | Average |
|---|---|---|---|---|---|---|---|---|
| Unsupervised | ACA | 84.5 | 92.5 | 60.0 | 92.2 | 87.8 | **92.8** | 85.0 |
| | HACA | 85.3 | 90.0 | 62.9 | 91.7 | 84.5 | 87.8 | 83.7 |
| | SC | 69.8 | 63.1 | 50.9 | 67.1 | 57.7 | 64.9 | 62.3 |
| | HDP-VAR(I) | 46.5 | 44.1 | 45.6 | 83.2 | **93.2** | 88.6 | 66.9 |
| Weakly Supervised | HDP-VAR(II) | 65.9 | 88.5 | 79.2 | 86.9 | 92.3 | 89.1 | 83.7 |
| Supervised | HDP-SLDS | 74.0 | 86.1 | 81.3 | **93.4** | 90.2 | 90.4 | 85.9 |
| | PS-SLDS | 75.9 | 92.4 | 83.1 | **93.4** | 90.4 | 91.0 | 87.7 |
| | XoN-MML | **85.3** | **93.0** | **87.0** | 90.6 | 88.9 | 88.2 | **88.8** |

*Table 3: Classification Accuracy Comparison with Other Techniques, all numbers in percentage. Bold fonts stand for the best accuracy on predicting the full sequences among methods. The last row includes the accuracies of the XoN-MML approach.*

Figure 8(a) illustrates the MML cluster pre-processing utilised in the No.2 honeybee dancing dataset model as an example. Figure 8(a) illustrates the common bee motion states as eight MML clusters, where the red arrows represent the movements from the previous frame, contrastingly, the blue arrows represent the current head direction. Figure 8(b) also includes histograms that indicate how the signatures or profiles of these eight MML cluster (Events) associate with the three behaviour categories.

A number of key facets can be understood in Figure 8(b): Firstly, the MML cluster events are a biased distribution across the three behaviours, which indicates the possibility to identify the honeybee behaviours based on these MML events. Secondly, the distribution histograms of "left-turn" and "right-turn" are similar to each other, therefore judging the type of behaviour with a single cluster/event is not feasible. For example, when new motion data is identified as cluster 0, it has little chance of it being a "Waggle", however the chance for being either a "Left-turn" or "Right-turn" are both very high. Considering that the final XoN model provides an accuracy of some 93% on classifying Sequence-2, this provides evidence of the effectiveness of this extended decision tree algorithm.



(a)

(b)

*Figure 8: (a) Presentation of MML clusters and (b) the Histogram of clusters on each behaviour. The very similar histograms of left- and right-turn indicate that the individual MML clusters are not the distinguishing features for these motion behaviours, therefore, the reason behind the successful classification lies in utilising the localised feature combination with the EGBC method.*

## 4.3 Hot Metal Temperature Prediction

As stated in Section 1, this research was motivated by a demanding industrial problem, i.e. to utilise data mining techniques to model and predict the iron-making process. In this section, the EGBC framework is employed for making predictions of the HMT, in order to evaluate how this method performs in a stochastic industrial environment.

### 4.3.1 Problem Definition

The HMT temperature of molten iron, is one of the most important indicators of product quality in the iron making process. However, due to the complexity of the associated non-linear processes and systems, predicting and stabilising this temperature is extremely difficult using standard statistical techniques or even mass-transfer equations (Biswas 1981).

The HMT is recognised as one of the most important indicators because it is directly associated with the product quality (Banks 1999), which is expected to be stabilised at an ideal value of 1,500°C with a tolerable ±20°C variance. Therefore, three classes, "High", "Norm" and "Low", are used to label every 'cast' (the process of tapping the molten iron) in the BF operational data.

The BF is continuously monitored by various sensors that are installed all over its structure. These sensors report a variety of operational data, such as the wall temperatures (via a complete array of water cooled blocks or staves), internal pressures near the wall, the contents of exhaust gases at the top, the blast temperature, plus the volume and humidity etc. Other information such as the rate of fuel, the volume or mass of raw feed materials being charged, laboratory test results and real-time simulation based indices are also collected as part of the monitoring log.

The data was collected regularly from each source, and the complete BF operational data was finally logged as a multi-dimensional time series. As the EGBC framework is employed here for predicting the possible changes of future HMT based on current and historical observations, the associated parameter settings (similar from the previous sections) will not be repeated in details in the remainder of this section.

### 4.3.2 Data Cleansing and Event Generation

The attributes in the original BF dataset were collected at differing sampling rates, some manually measured attributes are also entered at irregular time intervals. Within many data mining methods, this is often considered as a missing data scenario for the low frequency attributes. However, within the EGBC framework, events are constructed based on individual attributes, according to the temporal characteristic of each attribute, customised methods can be designed for feature generation from attributes with specific constrains, therefore the interpolation is not necessary on the original dataset because of this flexibility.

Although the temporal misalignment among attributes is inevitable, the EGBC framework is designed to accommodate such attributes with known time shifts, these are approximately aligned with other attributes in order to reduce the number of irrelevant events being included. By doing so, the grouping window for this task was consequently set at three hours, instead of more than eight hours as in the raw data.

Due to a lack of specific knowledge on all of the individual attributes, events in this case were selected to be the intuitively shape-based. An extended Symbolic Aggregate ApproXimation (SAX) algorithm, Variance-wise SAX (Sun et al. 2013), is utilised for dynamically converting all attributes in to SAX motifs. Within every training period, the mean and standard deviation values are recalculated based on the up-to-date history so that the equiprobable zones reflect the true historical information. Variance-wise SAX produces varying numbers of motifs within a selected period. Active attributes are automatically transformed into more SAX motifs with finer details of the sequence, and on the contrary, less active attributes incur fewer motifs that abstract larger time spans with fewer details.

The SAX motifs are further converted into events using a k-medoids clustering algorithm. In each training period, all SAX motifs from each individual attribute are gathered and clustered into a fixed number of clusters, and these clusters are in-turn used as the shape-based events. Because the number of event combinations generated during the XoN training process is exponential to the number of attributes and the number of individual features per attribute, and the capacity and performance of the test platform and the number of attributes in the BF dataset (over 30), the number of clusters for each attribute is limited to seven or nine clusters per attribute, such that the overall number of events are under control in the following XoN modelling stage.

### 4.3.3 Ensemble Learning with XoN Models

The XoN decision tree classifier becomes the central modelling function. However, because of the stochastic nature of the BF, it is unlikely that any individual model would fully describe the problem and produce reliable predictions. In order to improve the overall performance, multiple best performing base-XoN models are employed for building an ensemble model to vote for a final prediction of HMT in the test dataset. Additionally, as the process modality of the BF changes over time, models are retrained periodically by sliding through the whole ten months of data (in the step of 7 days) and segmenting these into 36 overlapped sets, each includes 40 days of training data and subsequently utilising the following 7 days as test data.

From each of these data segments, 48 XoN trees are generated by adjusting such parameters as: groupWin, clusterNo etc. Every tree is further expanded into 15 variations by altering the pruning parameters. Seven of the best performing historical models, from a prior (TS) segment plus two newly trained models form a pool of models in each new data segment – subsequent predictions are then made over the test data (remaining 7 days of a data segment) for evaluation. All XoN trees were fully grown and pruned back differently, as the following (subsequent) testing data is not available before the training finished, the newly added models are then selected based on evaluations over the training dataset, thus over-fitting is expected. However, over-fitted models can also be useful for ensemble models (Sollich and Krogh 1996). Further, all other models were evaluated and selected (and deselected) based on the historical testing data, therefore the possible occurrence of over-fitting during the training process is controlled and contained.

### 4.3.4 Outcomes and Comparison

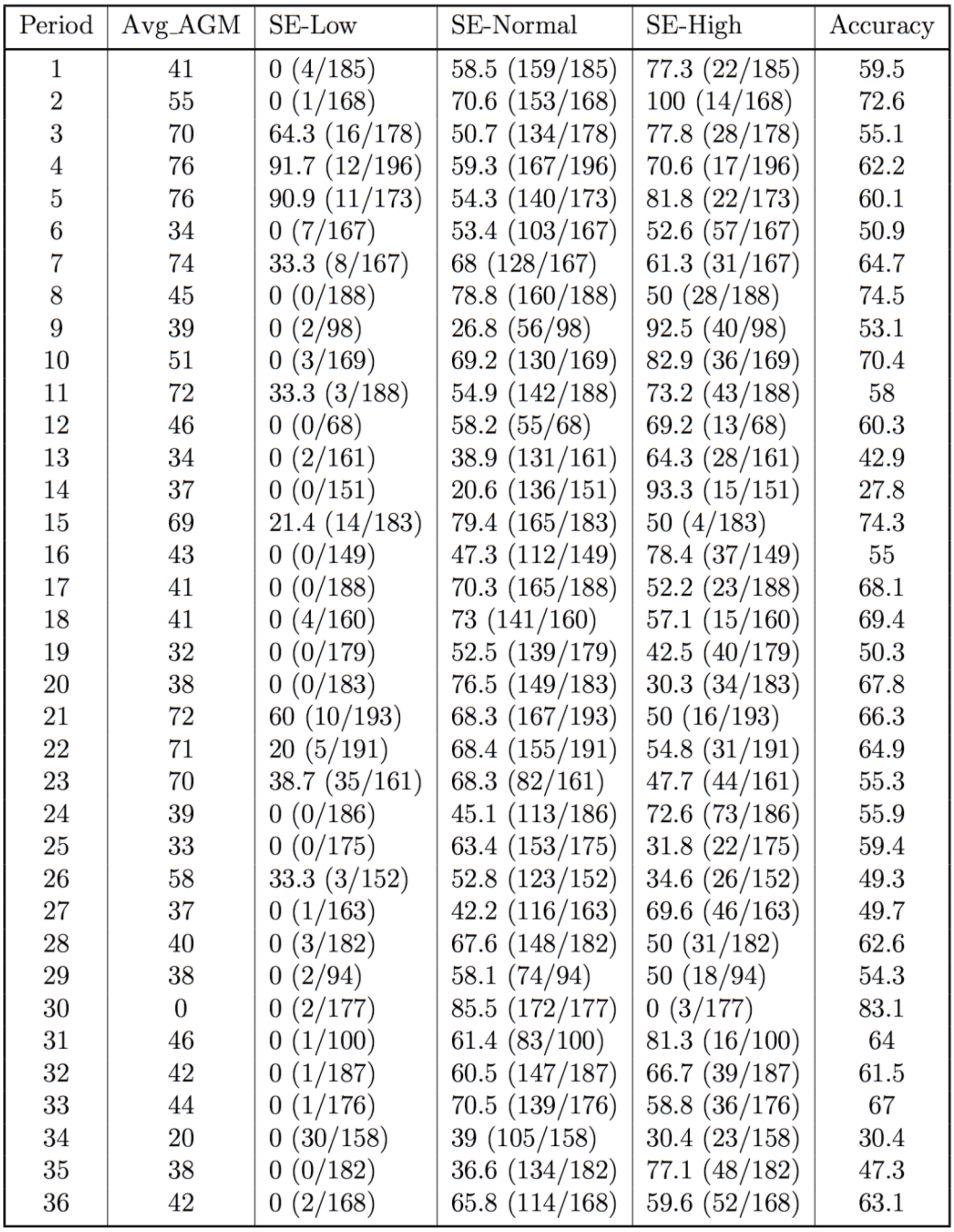
Because the BF data is an imbalanced 3-class dataset, in order to objectively evaluate the performance, the Adjusted Geometric Mean (AGM) (Batuwita & Palade 2012) is selected as the main metric for evaluating the prediction performance. The AGM is an extension of the GM (geometric mean) metric, which in turn is calculated from the four basic statistical measures (True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN)) with the following formulas:

$$AGM = \begin{cases} \dfrac{GM + TN_{rate} \times (FP + TN)}{1 + FP + TN}, & if\ TP_{rate} > 0 \\ 0, \quad if\ TP_{rate} = 0 \end{cases}$$

$$Where, \quad \begin{aligned} TN_{rate} &= \frac{TN}{FP + TN} \\ TP_{rate} &= \frac{TP}{TP + FN} \\ GM &= \sqrt{TP_{rate} \times TN_{rate}} \end{aligned}$$

The average AGM of the two abnormal HMT classes are listed in Table 4, together with a number of other performance indicators such as the true positive rates (Sensitivity) for both ‘Low’ and ‘High’ classes, and the overall prediction accuracy in all BF test periods. In general, for the Low-class there are less hits compared to the other classes, and the sensitivities of Low-class are often zero or other small values. However, because the Low-class cases are often rare in the various test datasets, this lack of hits on several occasions amongst some two hundred cases is somewhat understandable.

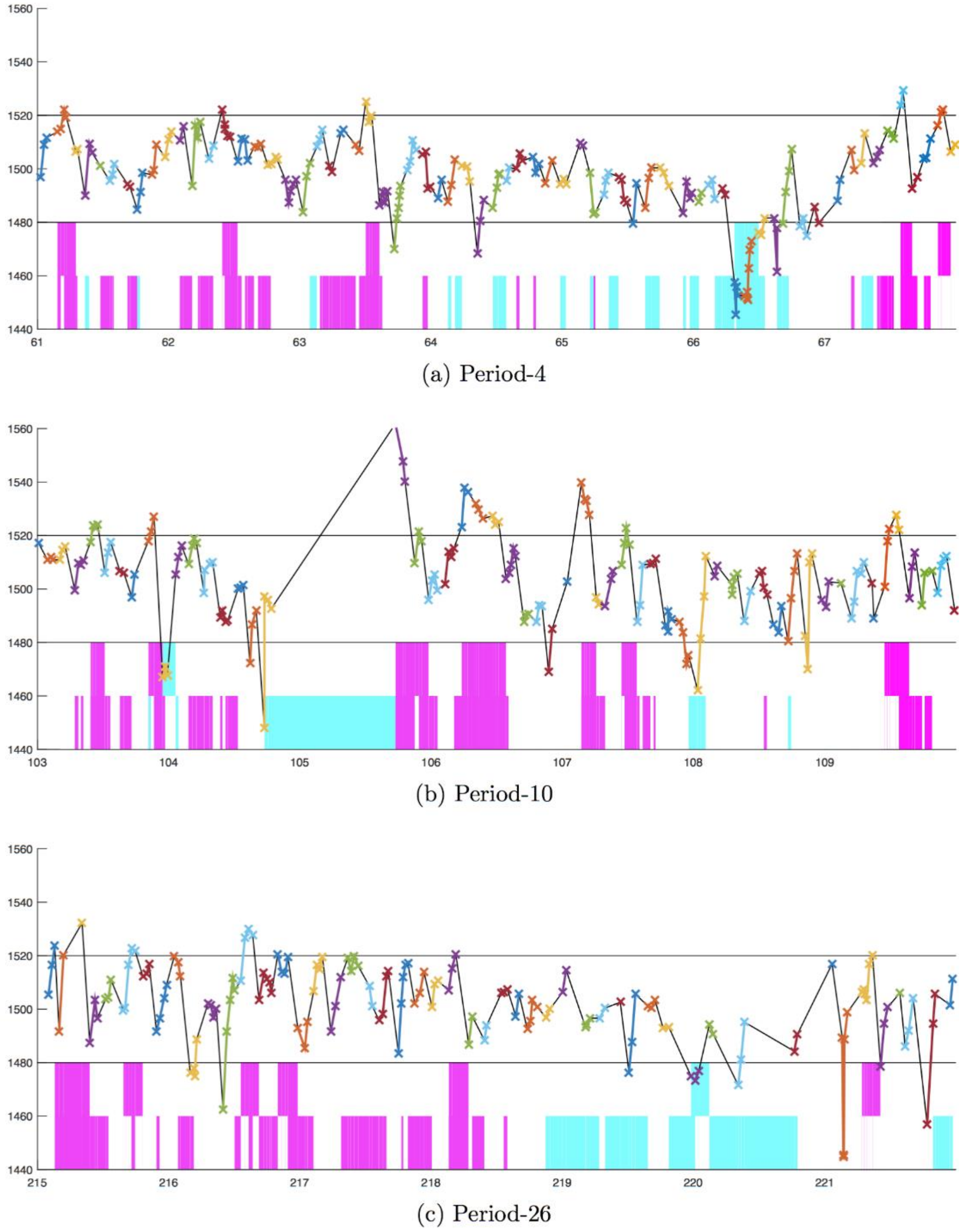
In the total 36 periods, the average test data size is approximately 170 cases per period. However, within these, there are some 26 periods that contain five or less instances of the Low-class, and the sensitivities of Low-class in these periods are all zero except Period-22. On the other hand, in the seven periods with more than ten Low-class cases, the Low-class sensitivities are 63.4, 91.7, 90.9, 21.4, 60, 38.7 and 0 (%) respectively. As for the other abnormal HMT class (High), the XoN tree models successfully predicted more than 60% of these High-class cases in half of the test periods, and only seven periods have sensitivities of less than 50%.



| Period | Avg_AGM | SE-Low | SE-Normal | SE-High | Accuracy |
|---|---|---|---|---|---|
| 1 | 41 | 0 (4/185) | 58.5 (159/185) | 77.3 (22/185) | 59.5 |
| 2 | 55 | 0 (1/168) | 70.6 (153/168) | 100 (14/168) | 72.6 |
| 3 | 70 | 64.3 (16/178) | 50.7 (134/178) | 77.8 (28/178) | 55.1 |
| 4 | 76 | 91.7 (12/196) | 59.3 (167/196) | 70.6 (17/196) | 62.2 |
| 5 | 76 | 90.9 (11/173) | 54.3 (140/173) | 81.8 (22/173) | 60.1 |
| 6 | 34 | 0 (7/167) | 53.4 (103/167) | 52.6 (57/167) | 50.9 |
| 7 | 74 | 33.3 (8/167) | 68 (128/167) | 61.3 (31/167) | 64.7 |
| 8 | 45 | 0 (0/188) | 78.8 (160/188) | 50 (28/188) | 74.5 |
| 9 | 39 | 0 (2/98) | 26.8 (56/98) | 92.5 (40/98) | 53.1 |
| 10 | 51 | 0 (3/169) | 69.2 (130/169) | 82.9 (36/169) | 70.4 |
| 11 | 72 | 33.3 (3/188) | 54.9 (142/188) | 73.2 (43/188) | 58 |
| 12 | 46 | 0 (0/68) | 58.2 (55/68) | 69.2 (13/68) | 60.3 |
| 13 | 34 | 0 (2/161) | 38.9 (131/161) | 64.3 (28/161) | 42.9 |
| 14 | 37 | 0 (0/151) | 20.6 (136/151) | 93.3 (15/151) | 27.8 |
| 15 | 69 | 21.4 (14/183) | 79.4 (165/183) | 50 (4/183) | 74.3 |
| 16 | 43 | 0 (0/149) | 47.3 (112/149) | 78.4 (37/149) | 55 |
| 17 | 41 | 0 (0/188) | 70.3 (165/188) | 52.2 (23/188) | 68.1 |
| 18 | 41 | 0 (4/160) | 73 (141/160) | 57.1 (15/160) | 69.4 |
| 19 | 32 | 0 (0/179) | 52.5 (139/179) | 42.5 (40/179) | 50.3 |
| 20 | 38 | 0 (0/183) | 76.5 (149/183) | 30.3 (34/183) | 67.8 |
| 21 | 72 | 60 (10/193) | 68.3 (167/193) | 50 (16/193) | 66.3 |
| 22 | 71 | 20 (5/191) | 68.4 (155/191) | 54.8 (31/191) | 64.9 |
| 23 | 70 | 38.7 (35/161) | 68.3 (82/161) | 47.7 (44/161) | 55.3 |
| 24 | 39 | 0 (0/186) | 45.1 (113/186) | 72.6 (73/186) | 55.9 |
| 25 | 33 | 0 (0/175) | 63.4 (153/175) | 31.8 (22/175) | 59.4 |
| 26 | 58 | 33.3 (3/152) | 52.8 (123/152) | 34.6 (26/152) | 49.3 |
| 27 | 37 | 0 (1/163) | 42.2 (116/163) | 69.6 (46/163) | 49.7 |
| 28 | 40 | 0 (3/182) | 67.6 (148/182) | 50 (31/182) | 62.6 |
| 29 | 38 | 0 (2/94) | 58.1 (74/94) | 50 (18/94) | 54.3 |
| 30 | 0 | 0 (2/177) | 85.5 (172/177) | 0 (3/177) | 83.1 |
| 31 | 46 | 0 (1/100) | 61.4 (83/100) | 81.3 (16/100) | 64 |
| 32 | 42 | 0 (1/187) | 60.5 (147/187) | 66.7 (39/187) | 61.5 |
| 33 | 44 | 0 (1/176) | 70.5 (139/176) | 58.8 (36/176) | 67 |
| 34 | 20 | 0 (30/158) | 39 (105/158) | 30.4 (23/158) | 30.4 |
| 35 | 38 | 0 (0/182) | 36.6 (134/182) | 77.1 (48/182) | 47.3 |
| 36 | 42 | 0 (2/168) | 65.8 (114/168) | 59.6 (52/168) | 63.1 |

*Table 4: Summarised HMT prediction performances of the EGBC framework over all test periods. The Avg AGM column shows the average AGM (adjusted geometric mean of the true cases) on both abnormal classes (Low and High), which is selected as the indicator of the prediction performance. Columns SE-* list the sensitivities of the predicting outcomes on all three classes, followed by the number of cases belonging to the given class and the number of total cases in the corresponding testing period. The last column shows the overall predicting accuracy on the test datasets, which is less significant than the sensitivities on abnormal classes.*

In addition to the prediction summary table, the prediction outcomes are visualised and three examples are given as in Figure 9. By examination of these visualisations, it is obvious that the EGBC models raise a number of false alarms. However, when and before a real abnormal HMT class happens, the ensemble model generally corrects this within a small-time proximity, thus such visual results may still be useful for BF control in reality.



(a) Period-4

(b) Period-10

(c) Period-26

*Figure 9: A few visual examples of HMT predictions at Period 4, 10 and 26. The bars at the bottom of pictures indicate real and predicted classes. Upper bars stand for real HMT classes, and lower bars are the outcome of predictions. Pink colour means the class is High and cyan colour means Low class. The line plot in the middle represent the measured HMT records, where each valid HMT sample is marked with 'x'. Different colours indicate different casts from the BF in a test period. X-axis represents days; Y-axis represents temperatures (°C).*

By comparing the prediction of EGBC framework to a previous study (Sun et al. 2011) on the same BF dataset employing C5.0 and Cubist decision tree models, the EGBC framework shows significant improvements. Although the overall prediction accuracies of the C5.0 approach are occasionally comparable or even better in certain cases than the EGBC approach, it is achieved by sacrificing the success or 'hit' rate for both abnormal HMT classes. In most of the data segments, the C5.0 predictions have almost zero sensitivity on the "Low" class, and the hit rate on the "High" class is also much poorer.

# 5 Conclusion and Future Work

This paper presents a novel method to utilise an XoN decision tree technique for classifying generic streams of data sequences. Provided that numeric time series data can always be transformed into a sequence of nominal events with an appropriate abstraction method, the XoN approach can be readily expanded to other time series data mining areas. The language identification test indicates that the XoN based approach has similar or better performance than the traditional LID algorithm, especially when classification is made on isolated words. The success on the honeybee behaviour classification demonstrates that our approach has the capacity to model and classify genuine real-valued time series data, with appropriate transformation of temporal events, and its stable performance, which is significant when compared with other techniques. However, the temporal misalignment does not exist in the two tasks above. In order to evaluate the merit of event group classification, plus to validate that it actually works as expected when dealing with the misalignment issue, real-world industrial data encompassing severe time shifting problems was studied. Here, the outcomes indicate a significant improvement in correctly predicting abnormal process states or modalities.

There are a number of advantages for the methodology presented in this paper. Firstly, the method looks for causally related signs for the classification, which naturally allows time variances between various attributes of the input data. Secondly, the method is based on an abstracted layer (beyond numeric values), the differences in attributes' physical meanings are ignored on this layer. Thirdly, temporal information and local ordering of a single attribute can be embedded into individual events, depending on the event transformation, and the sliding window ensures decision are made over a controllable period. However most importantly, unordered event combination allows temporal variance within the period, which simplifies the feature space and provides tolerance to temporal misalignment among attributes.

It has been illustrated in this work that the event group based decision tree approach demonstrates significant capability for classifying the sequential data. This event group based method now has the potential to be used as a generic methodology for modelling more complex time series data with missing or misaligned variates. However, users need to obtain considerable knowledge of the target problem in order to appropriately select the techniques for event generation, and the parameters for the EGBC modelling procedure, e.g. the size of sliding window, duration of events and size of the XoN nodes.

We also find the event groups generated by XoN have certain similarities to the concept of topic modelling (Steyvers & Griffiths 2007), however these are derived from supervised learning based on the information gain rather than probabilities. Future research will include utilising topic modelling for minimising or scaling down the feature space, in order to extend this method to more challenging real world time series data mining with multi-dimensional forms and potentially complicated intra-sequential relationships.

## References:

Agrawal, R., Faloutsos, C. & Swami, A. N. (1993), Efficient similarity search in sequence databases, in 'FODO', pp. 69–84.

Allen, J. F. (1983), 'Maintaining knowledge about temporal intervals', Commun. ACM 26(11), 832–843.

Banks, B. S. (1999), Neural network based modeling and data mining of blast furnace operations, PhD thesis, Massachusetts Institute of Technology.

Bhattacharjee, S. D. & Das, A. (1999), 'Application of artificial intelligence in tata steel,' Tata Search, Tech. Rep.

Batuwita, R. & Palade, V. (2012), 'Adjusted geometric-mean: a novel performance measure for imbalanced bioinformatics datasets learning.', J. Bioinformatics and Computational Biology 10(4).

Batyrshin, I. Z. & Sheremetov, L. (2008), 'Perception-based approach to time series data mining', Appl. Soft Comput. 8(3), 1211–1221.

Biswas, A. (1981), Principles of Blast Furnace Ironmaking: Theory and Practice, Cootha. URL: https://books.google.com.au/books?id=TBdxQgAACAAJ

Cavnar, W. B., Trenkle, J. M. et al. (1994), 'N-gram-based text categorization', Ann Arbor MI 48113(2), 161–175.

Fox, E. B., Sudderth, E. B., Jordan, M. I. & Willsky, A. S. (2008), Nonparametric bayesian learning of switching linear dynamical systems, in 'NIPS', pp. 457–464.

Freksa, C. (1992), An Event Group Based Classification Framework for Multi-variate Sequential Data, 'Temporal reasoning based on semi-intervals', Artif. Intell. 54(1), 199–227.

Granger, C. W. J. (1969), 'Investigating causal relations by econometric models and cross-spectral methods', Econometrica 37, 424–438.

Granger, C. (1980), 'Testing for causality: A personal viewpoint Journal of Economic', Dynamics and Control, 1980, 2, 329-352.

Harvey, J.-P. & Gheribi, A. E. (2014), 'Process simulation and control optimization of a blast furnace using classical thermodynamics combined to a direct search algorithm,' Metallurgical and Materials Transactions B, vol. 45, no. 1, pp. 307–327, 2014.

Hornik, K., Mair, P., Rauch, J., Geiger, W., Buchta, C. & Feinerer, I. (2013), 'The textcat package for n-gram based text categorization in R', Journal of Statistical Software 52(6), 1–17.

Kommenda, M., Kronberger, G., Feilmayr, C., & Affenzeller, M. (2011), 'Data mining using unguided symbolic regression on a blast furnace dataset,' in Applications of Evolutionary Computation. Springer, 2011, pp. 274–283.

Koza, J.R. (1992), 'Genetic Programming: On the Programming of Computers by Means of Natural Selection,' MIT Press, Cambridge, MA, USA.

Laxman, S., Sastry, P. & Unnikrishnan, K. (2007), A fast algorithm for finding frequent episodes in event streams, in 'Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining', ACM, pp. 410–419.

Mannila, H., Toivonen, H. & Verkamo, A. I. (1995), Discovering frequent episodes in sequences extended abstract, in '1st Conference on Knowledge Discovery and Data Mining, Montreal, CA'.

Mannila, H., Toivonen, H. & Verkamo, A. I. (1997), 'Discovery of frequent episodes in event sequences', Data Mining and Knowledge Discovery 1(3), 259–289.

McNamee, P. (2005), 'Language identification: a solved problem suitable for undergraduate instruction', Journal of Computing Sciences in Colleges 20(3), 94–101.

Mohammad, Y. & Nishida, T. (2010), 'Mining Causal Relationships in Multidimensional Time Series', Smart Information and Knowledge Management: Advances, Challenges, and Critical Issues, Springer Berlin Heidelberg, 309-338

Möller-Levet, C. S., Klawonn, F., Cho, K.-H. & Wolkenhauer, O. (2003), Fuzzy clustering of short time-series and unevenly distributed sampling points, in 'IDA', pp. 330–340.

Morrill, J. P. (1998), 'Distributed recognition of patterns in time series data', Commun. ACM 41(5), 45–51.

Oh, S. M., Rehg, J. M., Balch, T. R. & Dellaert, F. (2008), 'Learning and inferring motion patterns using parametric segmental switching linear dynamic systems', International Journal of Computer Vision 77(1-3), 103–124.

Porter, M. F. (1980), 'An algorithm for suffix stripping', Program: electronic library and information systems 14(3), 130–137.

Project Gutenberg. (n.d.), http://www.gutenberg.org. (Accessed on 15/Aug/2013).

Qiu, H.; Liu, Y.; Subrahmanya, N. A. & Li, W. (2012), 'Granger Causality for Time-Series Anomaly Detection', Proceedings of the 2012 IEEE 12th International Conference on Data Mining, IEEE Computer Society, 1074-1079.

Roddick, J. F. & Mooney, C. (2005), 'Linear temporal sequences and their interpretation using midpoint relationships', IEEE Trans. Knowl. Data Eng. 17(1), 133–135.

Rumelhart, D. E., Hinton G. E., & Williams, R. J. (1988), 'Neurocomputing: Foundations of research,' J. A. Anderson and E. Rosenfeld, Eds. Cambridge, MA, USA: MIT Press, ch. Learning Repre- sentations by Back-propagating Errors, pp. 696–699.

Sarma, R. K. (2000), 'Neural network based prediction and input saliency determi- nation in a blast furnace environment,' Ph.D. dissertation, Massachusetts Institute of Technology, 2000.

Shieh, J. & Keogh, E. J. (2008), isax: indexing and mining terabyte sized time series, in 'KDD', pp. 623–631.

Sollich, P. and Krogh, A. (1996), 'Learning with ensembles: How overfitting can be useful', Advances in Neural Information Processing Systems, volume 8, pp. 190-196.

Steyvers, M. & Griffiths, T. (2007), 'Probabilistic topic models', Handbook of latent semantic analysis 427(7), 424–440.

Sun, Y.; Li, J.; Liu, J.; Chow, C.; Sun, B. & Wang, R. (2015), 'Using Causal Discovery for Feature Selection in Multivariate Numerical Time Series', Mach. Learn., Kluwer Academic Publishers, 101, 377-395.

Sun, C., Stirling, D., Ritz, C. & Sammut, C. (2013), 'Variance-wise segmentation for a temporal-adaptive sax', Journal of Research and Practice in Information Technology, SI.

Sun, C., Stirling, D., Wright, B., Zulli, P. & Ritz, C. (2011), 'A hot metal based on hybrid decision tree techniques.', International Journal of Technology, Knowledge & Society 7(2), 37 – 48.

Takcı, H. & Soğukpınar, İ. (2004), Centroid-based language identification using letter feature set, in 'Computational Linguistics and Intelligent Text Processing', Springer, pp. 640–648.

van Noord, G. (n.d.), 'Textcat', http://odur.let.rug.nl/vannoord/TextCat/index.html. (accessed on 10/Aug/2015).

Wallace, C. S. & Boulton, D. M. (1968), 'An information measure for classification', The Computer Journal 11(2), 185–194.

Waller, M. & Saxén H. (2000), 'Applying nonlinear time series methods to blast furnace tap variables,' in Presented at Cybernetics & Informatics Eurodays, Marianska, Czech Reublic.

Waller, M. & Saxén H. (2002), 'Ontheuseofladle-wiseanalysestopredicthotmetal silicon content,' in Proceedings of the 61th Ironmaking Conference, Nashville, TN, pp. 135–143.

Zheng, Z. (2000), 'Constructing x-of-n attributes for decision tree learning', Machine Learning 40(1), 35–75.

Zhou, F., la Torre, F. D. & Hodgins, J. K. (2013), 'Hierarchical aligned cluster analysis for temporal clustering of human motion', IEEE Trans. Pattern Anal. Mach. Intell. 35(3), 582–596.

Zhou, P., Yuan, M., Wang H. & Chai T. (2015), 'Data-Driven Dynamic Modeling for Prediction of Molten Iron Silicon Content Using ELM with Self-Feedback,' Mathematical Problems in Engineering, vol. 2015, Article ID 326160, 11 pages, 2015. doi:10.1155/2015/326160.