

# Knowledge discovery and predictive accuracy comparison of different classification algorithms for mould level fluctuation phenomenon in thin slab caster

Preetam Debasish Saha Roy<sup>1</sup> · Prabhat Kumar Tiwari<sup>1</sup>

Received: 12 January 2016 / Accepted: 23 June 2016 / Published online: 6 July 2016  
© Springer Science+Business Media New York 2016

**Abstract** Mould level fluctuation (MLF) is one of the main reasons for surface defects in continuously cast slabs. In these study first, large scale mould level fluctuations has been categorized in three different cases based on actual plant data. Moreover, theoretical formulation has been investigated to better understand the underlying physics of flow. Next, exploratory data analysis is used for preliminary investigation into the phenomenon based on actual plant data. Finally, different classification algorithms were used to classify non-mould level fluctuation cases from MLF cases for two different scenarios- one where all mould level fluctuation cases are considered and in another where only a particular case of mould level fluctuation is considered. Classification algorithm such as recursive partitioning, random forest etc. has been used to identify different casting parameters affecting the phenomenon of mould level fluctuation. 70% of the dataset used as training dataset and rest 30% as the testing dataset. Prediction accuracy of these different classification algorithms along with an ensemble model has been compared on a completely unseen test set. Ladle change operation and superheat temperature has been identified as process parameters influencing the phenomenon of large scale mould level fluctuations.

**Keywords** Continuous casting · Mould level fluctuation · Classification algorithm · Data mining · Random Forest

## Introduction

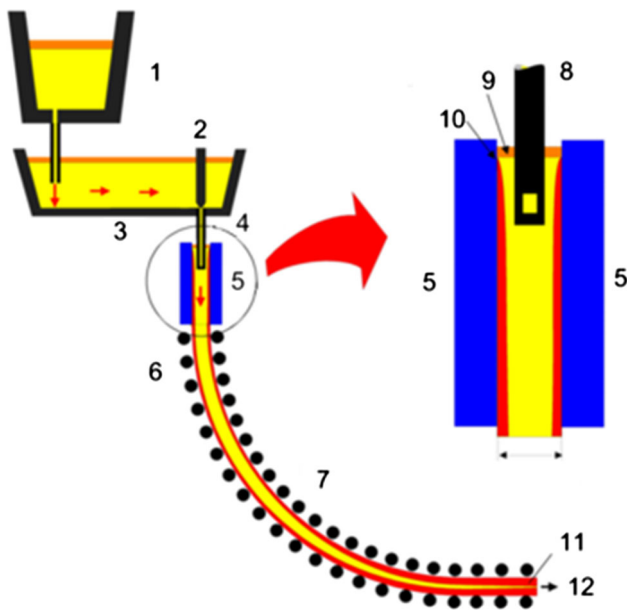
Continuous casting is the process by which molten metal is solidified into semi-finished slab, billet or blooms for subsequent rolling in the finishing mill. Previous to the introduction of the continuous casting technology after solidification slabs were stored in the yard and then taken to different location for the subsequent hot rolling. In the continuous casting process the whole process of casting, reheating and rolling process becomes continuous, thus increasing productivity greatly. In Tata Steel, Jamshedpur plant continuous casting technology is used for thin slab casting currently. Slabs of width ranging from 950–1700 mm and thickness of 57–72 mm are cast in the current thin slab caster plant.

Molten metal is first brought from ladle furnace using ladle and placed in the arm of the turret. Turret is used to alternate between the two ladles, after all the molten metal in one is drained to the tundish, next ladle is brought in its position. Molten metal is poured from the ladle to the tundish using a slide gate mechanism and shroud. Then from the tundish molten metal transported to the mould using a stopper-rod mechanism. Once the molten metal comes in contact with the water-cooled copper plate, then molten metal solidifies to form a thin shell around the mould. Drive rolls in the lower part continuously withdraws the solidified shell from the mould at a certain speed which is more or less equal to the rate of incoming hot metal. This withdrawal speed of the strand is known as the casting speed. Below the mould metal shell passes through five more segments and finally through the bending unit, before being cut into different slabs in the shearing machine. After exiting the mould the solidified shell supports the core liquid metal up to the metallurgical length, where the whole slab is solidified (Fig. 1).

Mould powder is added to the top of the surface of the mould to provide thermal and chemical insulation. These

✉ Preetam Debasish Saha Roy  
preetamsr@gmail.com

<sup>1</sup> Automation Division, Tata Steel, Jamshedpur 831005, India



**Fig. 1** Continuous casting process. 1 Ladle. 2 Stopper. 3 Tundish. 4 SEN. 5 Mould. 6 Roll support. 7 Turning zone. 8 SEN. 9 Bath level. 10 Meniscus. 11 Withdrawal unit. 12 Slab (Javurek 2008)

mould powder melts and floats on top of the free surface of molten steel in the mould. At the same time the mould is also oscillated continuously to facilitate stripping of the solidified metal shell from the mould wall and it also helps in uniform distribution of the mould flux.

Initial solidification of the molten metal starts at the meniscus and the behaviour of the meniscus also effects the heat transfer process down-stream, so the time dependent behaviour of the meniscus plays an important role in determining the final quality of the product. Transient fluctuations of the top free surface in the mould is known as mould level fluctuation (MLF). Sudden fluctuations in the mould level can be very detrimental as it interrupts the solidification process and it can entrap the mould flux in the molten steel, leading to surface defects in the final rolled product.

In literature, sudden fluctuations of the mould level is associated with the surface defects in the final product, but when these fluctuations are very large, then mould flux entrapment in the molten steel causes sudden change in width of the cast slabs. These in turn can cause cobble in the mill, due to sudden increase in the thickness of the slab fed to the rolling mill. In continuous casting operations as there is no intermediate point where these defective slabs can be diverted, so it may cause cobble in the mill, causing interruption in the entire production process. In thin slab caster (TSCR) facility this problem was identified as one of the reasons of cobble in mill. So, mould level fluctuations information captured from Caster level-1 automation system was used to assign MLF signal to a corresponding slab in caster level-2 system and these information was sent to the mill level-2 system prior to

the slab being rolled. Upon receiving the MLF signal, mill level-2 system increases the roll gap for smooth rolling operations, thus cobble can be avoided.

Objective of this study is to identify casting parameters affecting these large mould level fluctuations in the thin slab caster, so to better understand the phenomenon. In this study first exploratory data analysis was used to understand and identify different casting parameters that influence the phenomenon of mould level fluctuations. Also theoretical formulations about the fluid flow in the mould was investigated to understand the underlying physics of flow. Next, different classification algorithms were used to identify major process parameters influencing the phenomenon of mould level fluctuation in thin slab caster facility, Tata Steel, Jamshedpur plant. Classification algorithms based on 70 % of the data were tested on the completely unseen rest 30 % of the data to verify the predictive accuracy of the models. Moreover, predictive accuracy of these classification algorithms were compared along with stacked model where results of all the other classification algorithms were considered.

## Literature review

Due to the detrimental effect of mould level fluctuations on the final rolled product, extensive research work has been carried out academically and industrially to understand the phenomenon properly. Thomas (2001) in his review of different aspects of continuous casting process, also mentions the importance of shape of the top surface in determining the final quality of the steel. Understanding time variant shape of the top surface gives us better insight about the mould level fluctuation phenomenon. Lee et al. (2012) compared the metal level fluctuations phenomenon in the continuous casting process with the “butterfly effect”, as mould level fluctuations coupled with heat transfer process can cause oscillation marks, cracks, tears or even breakouts. They also emphasis on the limitation of the numerical models due to the inherent chaotic conditions inside the mould. Singh et al. (2011) also in their integrated approach of modelling continuous casting process, emphasis on the importance of the considering tundish and caster simultaneously, rather than the popular ‘silo’ based approach. Sometimes simultaneous fluid flow simulations (considering turbulence effects) and control system analysis is used for better control of the mould fluid level (Suzuki 2004), but control of the mould level was highly sensitive to the spatial arrangement of the sensors leading to inaccurate control in some areas. Such studies involving industrial applications underline the importance of this phenomenon in determining the final product quality.

Other authors utilised different experimental studies to understand the different defects generated by MLF. Shaver (2002) explored the possibility of measuring meniscus steel

velocity with inexpensive nail board method. In this study FFT analysis along with velocity measurements were utilized for better understanding of the defects generated due to flux entrapment. Thomas (2013) in his extensive cover of fluid flow phenomenon in the mould points out that severity of the defects in case of large scale fluctuations in mould level, which is of interest of present study.

In different studies upstream and downstream events have been explored extensively to correlate these events to the MLF phenomenon. Liu et al. (2012) used transient CFD simulation model with multiphase flow of steel and gas bubble in the sub-merged entry nozzle (SEN) and in the mould to study the effect of stopper rod movement on the mould fluid flow. Results suggest significant disturbances of the meniscus occur during the movement of the stopper rod, resulting in flux entrapment and formation of sliver defects in the final product. In another investigation by Liu et al. (2013) large metal level sloshing was observed when dithering frequency matched with the mould frequency. In these studies comparison of the numerical results with the plant data suggest some dissimilarities which may be attributed to the turbulent flow of liquid metal in the mould. Sudden flushing of the inclusion deposits also has been pointed out as possible reason for mould level fluctuations by Girase et al. (2007), mainly due to  $\text{Al}_2\text{O}_3$  inclusions. This nozzle clogging phenomenon also discussed in detail by Rackers and Thomas (1995). Apart from these upstream events, downstream events like slab bulging has been reported as one of the reasons for mould level fluctuations in different studies. Matsumiya (2006) suggested the case of unsteady slab bulging, in which liquid metal is pushed up and down as the bulged slab passes through different rollers in the downstream process. Furtmueller et al. (2005) suggested reduction of the roll gaps and reduction of casting speed as possible measures to tackle the problem of bulging. They also mention the case of severe mould level fluctuations (Mould level hunting) in a plant setting where casting has to be stopped. While these studies provide great insight into particular phenomenon, they do not encompass all the other factors that may give rise to non-linear behaviour.

Others have explored complex flow physics and heat transfer mechanism inside the mould using CFD and experimental studies. Liu et al. (2011) studied the effect of stopper rod movement on the transient flow phenomenon in the mould and formation of sliver defects using 3D computational fluid dynamic model. They also suggested mathematical model for mould level using SEN flow rate, cast speed and other casting parameters. Results of their simulations required a less diffusive turbulence model to capture small scale transience. Hajari and Meratian (2010) studied fluid flow features in a funnel shaped mould with tetra-furcated nozzle. Their experimental investigation suggested while gas injection aids in increasing casting speed, it also increases the surface tur-

bulence. Although their study considers a thin slab casters but the study is limited by the special nozzle design, which can significantly affect flow physics inside the mould. Zhang et al. (2007) suggested asymmetrical flow patterns during transient events such as change in casting speed, change in gas flow rate and ladle change as possible reasons for metal level fluctuations. Effect of higher casting speed and SEN submergence depth has also been investigated in their study. Their study also suggests metal slag inclusion increases by two fold during ladle change and it increases ten times during the start and end of a sequence.

Effect of different transient operations has been explored in detail by some studies and the ladle change operation has been related to fluid flow in mould in some of these studies. Kumar et al. (2007) developed an mould level fluctuation index in an operational plant and the plant data suggests deviation of the index from the optimum values during ladle change and width change due to asymmetry in the flow. Gurosoy (2014) investigates effect of different flow controllers on the mould fluid flow including during ladle change. Their study can be extended to consider effect of the transient fluid flow and unsteady turbulent flow. Kant et al. (2013) investigated the effect of dam position on intermixing in a six strand billet caster during ladle change over. Results of the numerical and experimental study suggests position of the dam in the tundish plays crucial role in determining the liquid metal flow from tundish to the mould.

In recent times different data mining techniques and artificial intelligence methods are being used to tackle the most complicated problems in the steel industry. Tang et al. (2005) propose a neural network method for hybrid flow shop scheduling problem with dynamic job arrival consideration. The neural network model consisted of three sub-network, each corresponding to three different stages of steel making, i.e. Steel making, refining and casting. Wang and Wu (2003) investigate a mixed integer programming model for the case of multi-period, multi-product, multi-resource production scheduling problem. Nastasi et al. (2016) compared three different genetic algorithms as applied to the multi-objective storage strategy of an automated warehouse in steel industry. They implement widely accepted statistical tests to establish the superior performance of the three genetic algorithms over the previously used heuristic procedure. Ordieres-Meré et al. (2008) discuss the performance of different linear and non-linear learning algorithms for mechanical property prediction of galvanized steel coils. Also, the effect of knowledge based segmentation of the steel grades on the predictive performance of this methods are also discussed. Sharafi and Esmaily (2010) used decision trees, neural networks, association rules to predict pit and blister defects in low carbon steel grades. de Beer and Craig (2008) used decision tree, fuzzy logic, statistical regression to investigate an industrial online model for width prediction. Colla et al. (2011) used deci-

sion tree and rectangular basis function networks to predict under pickling defects and suggest optimal speed or speed range for the process line. While their study focuses on the under pickling defects, incorporating over pickling defects in the same model may alter the performance of the algorithms considerably.

## Theoretical formulation

Liu et al. (2011) modelled SEN inlet flow rate into the mould based on measured mould level and casting speed. According to their mass-balance model,

$$Q_E(i) = \frac{h(i+1) - h(i-1)}{2\Delta t} \times \left( W \times T - \frac{\pi d_{SEN}^2}{4} \right) + V_{cast}(i) \times W \times T$$

where  $Q_E$  is the molten metal flow rate through the SEN,  $h$  is the metal level in the mould,  $W$  is the width,  $T$  is the thickness of the mould and  $V_{cast}$  is the casting speed, also the index  $i$  denotes the particular time step.

Now, this equation can be modified as,

$$Q_E(i) = \frac{h(i+1) - h(i)}{\Delta t} \times (A_m - A_{SEN}) + V_{cast}(i) \times A_m$$

where  $A_m = W \times T$  is the mould cross sectional area and  $A_{SEN} = \frac{\pi d_{SEN}^2}{4}$  is the area of the SEN nozzle. Now if we write this equation in a slightly different way then we will get,

$$h(i+1) = h(i) + \frac{[Q_E(i) - V_{cast}(i) \times A_m]}{(A_m - A_{SEN})} \Delta t$$

In this equation we further assume that the cross-sectional area of the mould is much greater than the SEN cross-sectional area, so

$$h(i+1) = h(i) + \left[ \frac{Q_E(i)}{A_m} - V_{cast}(i) \right] \Delta t$$

From this formulation we can estimate metal level in the mould and can get a better insight about the metal level variation. According to the previous equation from one time step to the next metal level in the mould increases by an amount equal to  $Q_E(i)/A_m - V_{cast}(i) \times \Delta t$ . So, this gives us the intuition that metal flow rate through the SEN, cross-sectional area of the mould and cast speed are three important parameters influencing the mould level fluctuation problem. The current practice in implementation doesn't differentiate between the rise or fall of the mould level, so we should consider the modulus of the term represented within parenthesis.

## Current methodology

Inclusions during the casting processes can be classified into two groups—indigenous and exogenous. Indigenous inclusions are by products of large scale steel making process, which cannot be avoided. Exogenous inclusions are introduced during molten metal transition from ladle furnace to the caster and during the casting process itself. These exogenous inclusions are more detrimental to the final quality of product. While most of the research focuses on the surface defects due to mould level fluctuations, large scale fluctuations cause another problem. Due to material inclusion cast slab thickness varies along the length of the slab, which may cause cobble in the mill if roll gap is low. So, mould level fluctuation information needs to be transferred to the mill level-2 system prior to the slab being rolled. After getting the information from the level-1 PLCs, in caster level-2 system this signal of mould level fluctuation was assigned to the particular slab and this information is sent to the mill level-2 system, so that mill cobble can be avoided.

Mould level is measured using a radioactive source. This mould level data is recorded every 0.001 s. Whenever the value of mould level fluctuates more than 10% of its value in the previous time step, it is recognized as a large scale mould level fluctuation, so MLF signal is generated. Suppose mould level is denoted by  $h$  and  $i$  denotes the time-step index, then whenever  $h(i+1) - h(i) > 0.1h(i)$ , then MLF signal is generated.

In Fig. 2, two cases have been compared, one where mould level fluctuation signal was generated and where no signal was generated. To quantify mould level fluctuation a variable *MLF* was created, where

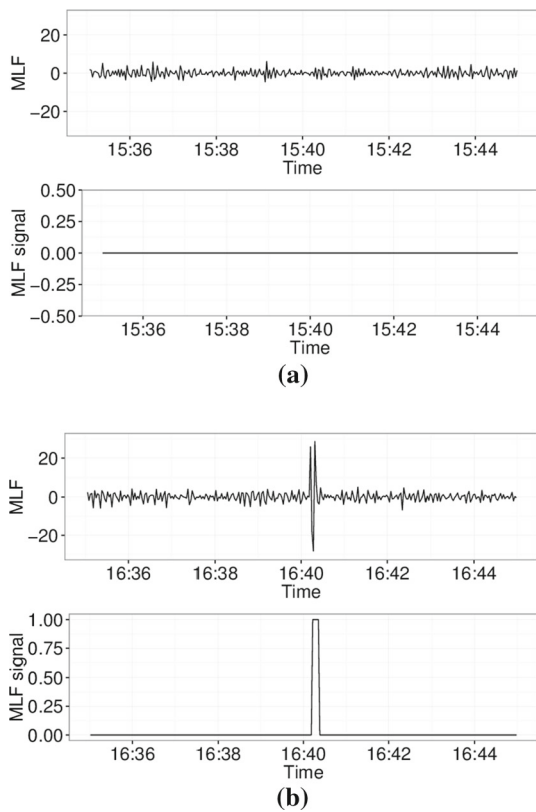
$$MLF = \frac{h(i+1) - h(i)}{h(i)} \times 100$$

Here  $h(i)$  denotes the value of the mould level captured in caster level-2 from caster level-1. In this communication data is captured every 2 s.

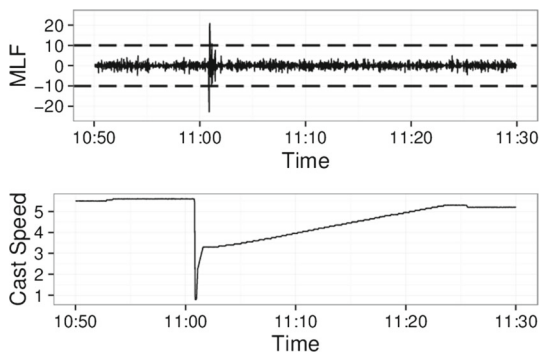
As we can see in Fig. 2a the *MLF* value never crosses the critical limit of 10% in the time window considered, so no MLF signal was generated, but in the second case when the value of MLF signal exceeds the threshold limit of 10% a MLF signal was generated.

Next, all the MLF slab data were collected from the mill level-2 system for last three months of 2014. Preliminary data analysis suggested relation between MMS (mould monitoring system/breakout prediction system) alarm and MLF slabs, as most of the MMS alarm cases were followed by a sudden change in mould level value. This is an expected result, because when a MMS alarm is generated, casting speed is automatically slowed down to a speed below 1 m/s to prevent imminent breakout.





**Fig. 2** Visualization of mould level fluctuation and MLF signal **a** case where no MLF signal was generated, **b** case where MLF signal was generated



**Fig. 3** Mould level fluctuation due to slow down after MMS alarm

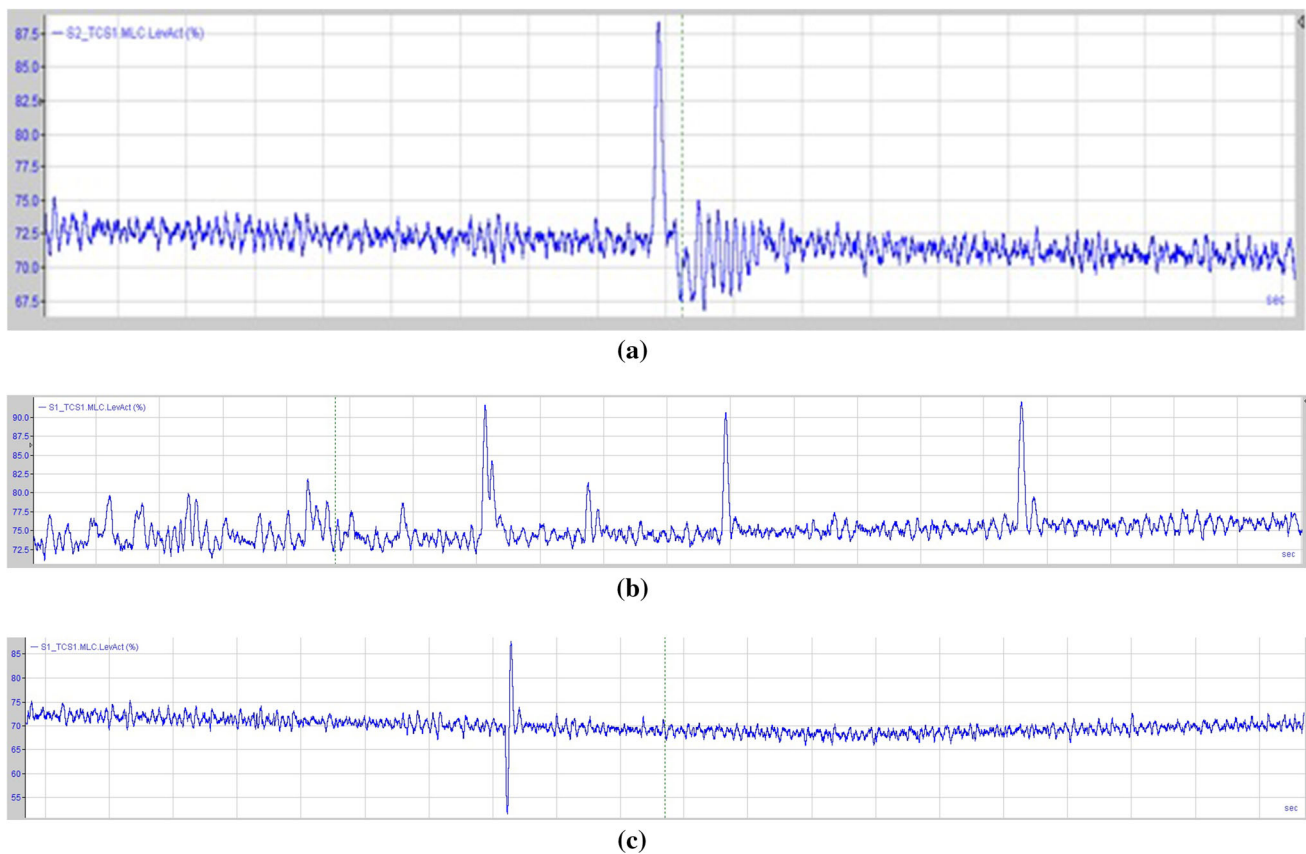
Also these intuition can be visualised using above mentioned mathematical formulation. When a MMS alarm is generated, casting speed is slowed down drastically, as we can see in the Fig. 3 the casting speed has been reduced to below 1 m/s from a value greater than 5 m/s. Due to this sudden decrease in the casting speed the term controlling metal level also changes drastically. During slowdown although casting speed is slowed down, the inlet steel flow rate is not controlled simultaneously, so it takes times for the mould level control mechanism to regain control over the metal level in the mould. So, almost all the cases of MMS alarm were followed by an MLF signal.

Another case in connection with mould level fluctuation needs to be considered is the lead heat case. From our analysis we can exclude lead heat cases safely, as lead heat slabs are not rolled for prime customers and also at the starting of a sequence roll gap is kept larger in the rolling mills. So our problem of sending the MLF signal and increasing roll gap is taken care of by default. Other than this two cases, all other cases of large scale mould level fluctuations are of particular interest for this study. Remaining MLF cases are denoted as true MLF cases in rest of this study. This true MLF cases are of our prime interest, as the cause of this large scale mould level fluctuations is not properly understood.

### MLF signal characteristic

Next mould level values from the level-1 system were analysed using signal analysis software IBA Analyser. In level-1 mould level data is sampled at a frequency of 0.001 s. While observing these data, interesting features were noticed. Particular type of MLF corresponded to a certain pattern of variation in mould level data. These patterns explained in Fig. 4. Let’s look at the three cases individually,

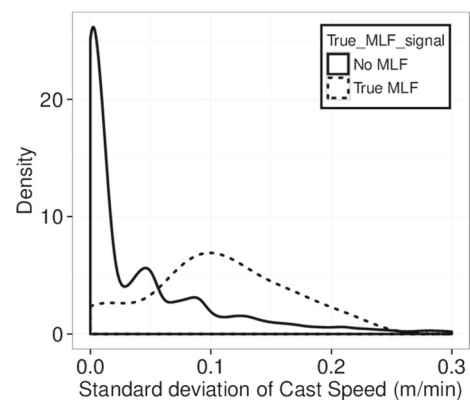
- (a) In Fig. 4a a particular MLF case which was due to slow down is shown. After a MMS alarm casting speed is slowed down drastically (generally from above 5 m/s to less than 1 m/s), but the inlet metal flow rate is not adjusted, so a sudden fluctuation in metal level is noticed. In this case it takes time for the mould level controller to gain stable control over the mould level. So, after a big fluctuation some small fluctuations are noticed, before the mould level becomes more or less stable. Also, in this case due to slowdown while the outflow from the mould is reduced but inflow is not reduced at the same manner, so in general a sudden rise in the mould level is observed.
- (b) In Fig. 4b a particular MLF case which was a lead heat (first heat of the sequence) case is shown. During the starting of a sequence dummy bar is inserted in the mould to support the liquid metal initially and it is slowly withdrawn. During lead heat, due to initial turbulence of the fluid flow in the mould lot of large scale fluctuations were noticed.
- (c) In Fig. 4c a particular MLF case which is denoted as true MLF case (up till now in this study) is shown. In this particular cases of MLF only single large scale fluctuations are noticed. Most of the true MLF cases analysed show similar trend, indicating a particular reason for all the cases. In all the cases a sudden drop in mould level was observed, followed by a sudden jump in metal level.



**Fig. 4** Characteristic of mould level values for three different cases—**a** MLF signal after MMS alarm case, **b** Lead heat MLF case and **c** True MLF case

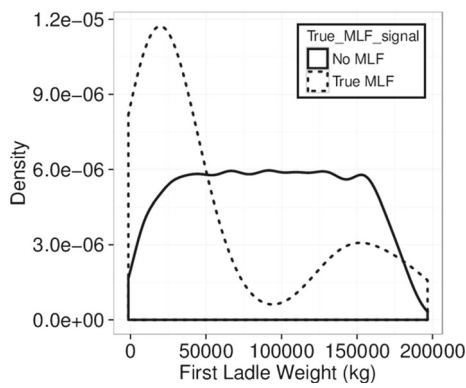
### Exploratory data analysis

In order to investigate the cause of large scale mould level fluctuations for the true MLF cases, exploratory data analysis was used. This mould level fluctuation can be caused by different events such as nozzle clogging (Girase et al. 2007), unsteady bulging (Furtmueller et al. 2005) etc. Reasons for deviation of mould level from the set-point values can be divided in two categories mainly, (a) mechanical transience from downstream and (b) fluid flow effects from upstream. Mechanical transience includes slab shear cut, unsteady bulging, oscillation effects etc. and flow related fluctuations include nozzle clogging, nozzle submergence, flow turbulence etc. Data for the month of January, 2015 was collected from data warehousing system IPQS for the exploratory analysis. During data collection it was noticed almost all the true MLF cases were in penultimate slab of the heat, only in one cases it was the last slab of the heat. Further exploration of the data suggested strong linkage of the true MLF cases with the ladle change operation. So for the exploratory analysis focus was given to upstream parameters that may affect the mould level fluctuation.

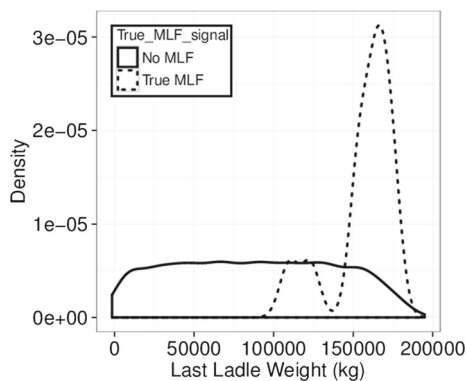


**Fig. 5** Comparison of standard deviation of the casting speed for the cases of no MLF and true MLF

For the exploratory data analysis part 41 parameters were considered like argon flow, ladle weight, casting speed, C% in tundish, superheat temperature, tundish weight etc. Out of the 41 process parameters lot of the variables were minimum, maximum and standard deviation of the same parameter during casting of a particular slab. Now probability density plot of these variables were plotted for both cases where no MLF



**Fig. 6** Comparison of first ladle weight for no MLF and true MLF cases



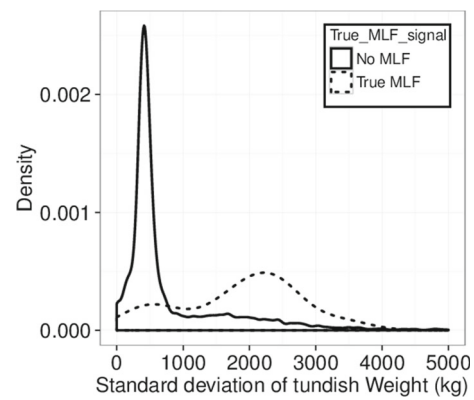
**Fig. 7** Comparison of last ladle weight for no MLF and true MLF cases

signal was generated and where true MLF signal was generated. For this part of analysis the other two cases of MLF were removed from the dataset as part of pre-processing to solely focus on the true MLF cases.

Out of all the variables in some of the variables appreciable difference was noticed between true MLF and non-MLF cases. Although the data set is highly imbalanced in favour of Non-MLF cases. Some of these are explained in the following section.

In Fig. 5 standard deviation of the casting speed has been compared for the cases where no MLF signal were generated and true MLF were generated. For non-MLF cases standard deviation of the casting speed is much smaller when compared against true MLF cases. So, it gives us the intuition that in case of true MLF there seems to be larger variation in casting speed for a particular slab.

Next, in Figs. 6 and 7 ladle weight has been compared for both the cases where non-MLF were generated and where true MLF signal was generated. In particular two variables were chosen from the IPQS- the first ladle weight, which corresponds to the ladle weight at the start of the casting of the particular slab and the last ladle weight, which corresponds to the ladle weight at the end of the casting of a particular



**Fig. 8** Comparison of standard deviation of the tundish weight for no MLF and true MLF cases

slab. From this again we can corroborate the claim that all the true MLF cases are generated during ladle change.

In Fig. 8, standard deviation of the tundish weight has been compared for both the cases where no MLF signal was generated and where true MLF signal was generated. Figure 8 suggests a larger variation in tundish weight for true MLF cases when compared against the non MLF cases. This is also in accordance with our finding that true MLF cases were generated only during ladle change. During ladle change, there is no incoming metal flow into the tundish, so tundish weight decreases sharply and this change is also reflected in the Fig. 8.

To conclude our results from the exploratory data analysis we can say that all the true MLF cases were generated during ladle change. This finding is also in accordance with literature where Zhang et al. (2007) also found increase in metal slag inclusion during ladle change. Also, Kumar et al. 2007 reported deviation of the mould level fluctuation index from optimum value during ladle change for operational plant. So, these reported cases of mould level fluctuation or slag entrapment defects during ladle change operation gives us confidence in our finding. Other than this, for true MLF cases the variation of casting speed was found to be more when compared against non-MLF cases.

### Classification approach

In this section we tackle the problem of large scale mould level fluctuations using different classification algorithms, commonly used in data mining. For this section of the study and the plots generated in the preceding section open-source statistical computing software R (R Core Team 2014) has been used.

For this part of the analysis first data were collected from the IPQS for the month of January, 2015. In total 41 variables were considered. Due to the transient nature of the

continuous casting process, for some of the attributes variance with time is very important. So, to capture this inherent temporal nature of the phenomenon minimum, maximum, average and standard deviation over the period of casting of one slab was considered for some of these variables. These kind of variables included ladle shroud argon pressure, ladle weight, casting speed, mould level, mass flow rate through SEN (submerged entry nozzle), tundish weight, argon flow rate, superheat temperature etc. Apart from these variables two binary categorical variables like thickness change in progress and width change in progress were also considered. Other variables included C% in the steel, slab thickness and width. As a part of data cleaning process NA (not available) values were removed from the data set. As the standard deviation of the ladle shroud argon flow rate contained almost half of its values as NA, so it was removed from the dataset. After data cleaning our data set contained 9437 observations (each observation is a slab only) and total 40 predictor variables. For this initial part of the study no distinction was made between the different types of MLF.

Out of three different types of MLF cases, MLF after MMS alarm cases and lead heat MLF cases are unavoidable due to dependence on other systems and current operational practices respectively. The true MLF cases are of primary interest for this analysis, as assignable causes for these type of cases are not fully understood. So to gain insight about the true MLF cases from the earlier mentioned dataset other two cases of MLF were excluded. Now the data set contains 9381 observations and 40 predictors consisting of only no MLF and true MLF cases. Now this data set becomes even more highly imbalanced, as there were only 12 true MLF cases out of this 9381 observations. Number of positive cases is less than 1%. So, for this part of this analysis, synthetic minority oversampling technique (SMOTE) algorithm (Torgo 2010) was used on the data set to make it more balanced. The SMOTE algorithm was set up for five time over-sampling of the true MLF cases and for each over-sampling five nearest data points were utilized. Also, for each over-sampling case five No MLF cases were randomly selected to complete the dataset. After the applying SMOTE algorithm, in data set 20% of the cases were true MLF cases.

As a part of the preliminary analysis, the dataset was explored to find patterns in the data that may explain anomalous cases of true MLF. First hierarchical clustering of the dataset was performed with Euclidean distance measures. In Fig. 9a, results of this clustering has been visualized as a dendrogram (Galili 2015) against the class labels. True MLF cases are represented as black against the grey No MLF cases. Results suggest that the No MLF and true MLF cases are distributed over the entire range and true MLF cases are not impacted by a particular subfamily. Similarly, classical multi-dimensional scaling was performed on the dataset to reveal underlying patterns in data structure. Multi-dimensional scal-

ing tries to preserve the between-object distances as well as possible. As a pre-processing stage, all the attributes (except two variables) of the data set was scaled to a mean value of zero and standard deviation of 1. Two variables, i.e. thickness change in progress and width change in progress were considered as binary categorical variables and no scaling was performed. Figure 9b visualizes the results of this multi-dimensional scaling in two dimensions. Although, some of the true MLF cases are clustered together, but the rest of the cases are not easily separable from the no MLF cases.

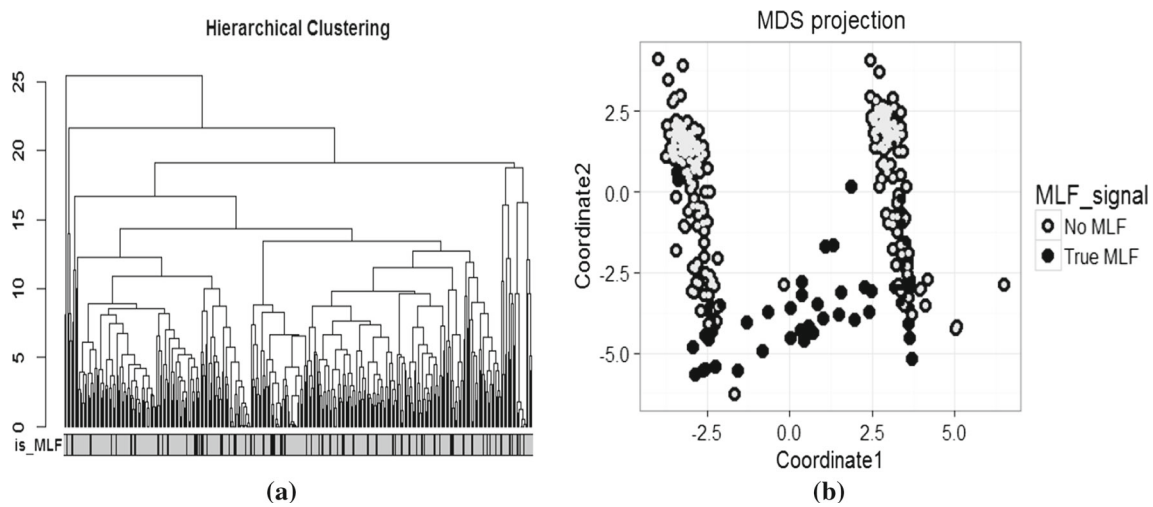
Along the same line Self-organizing mapping (SOM) (Wehrens and Buydens 2007) was also explored to identify hidden features in the data. Unlike multi-dimensional scaling which focuses on distance between data points, SOM focuses on preserving the topology in lower dimensions. So, objects that are close in two dimensional SOM projection are very similar in multi-dimensional space as well. Results of SOM are visualized in Fig. 10, where the left hand image shows the count of data points in each node, right side image maps these data points based on their class labels. Again, no clear pattern emerges to distinguish true MLF cases from no MLF cases.

In this section the training methodology for different classification algorithms are discussed that were utilized in this study. The whole dataset was divided in two different groups-training data and testing data. Training data contained 70% of the total data, while testing data contained the rest. This training data was used to train our dataset to different learning algorithms and results were validated using the test dataset.

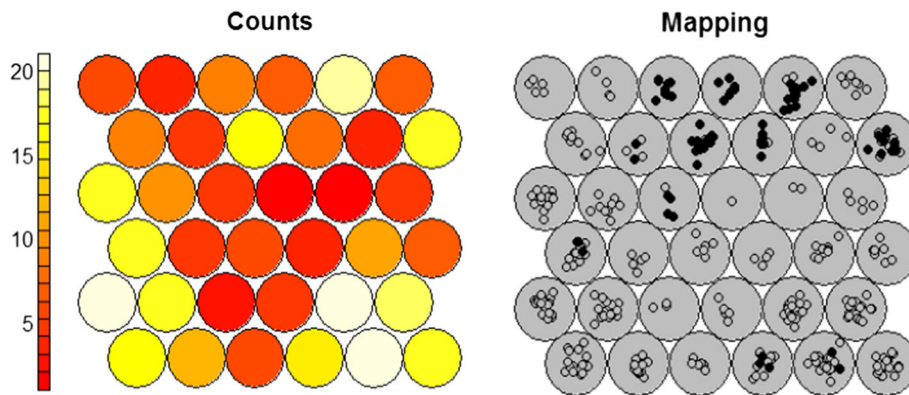
### (a) rpart

First the rpart model is utilized with minimum bucket size of zero and zero value for the complexity parameter. The minimum bucket size denotes the minimum number of cases in the leaf nodes and complexity parameter is representative of the improvement of the accuracy of model at each split at the cost of complexity. Also, tenfold cross-validation was implemented on the training data set to train the rpart algorithm. By implementing zero complexity parameter and zero minimum bucket size complex tree is formed, which highly prone to over-fitting the training data and thus performing poorly in any unseen data. So, tree pruning is required for the model to generalize well over whole data set. In Fig. 11, cross-validation relative error on the training data set is plotted against the tree size. As, the tree grows error decreases initially, but after certain extent due to over-fitting the model performs poorly on the validation set and thus error increases. So, for tree pruning complexity parameter corresponding to the lowest error rate was utilized for final model building.

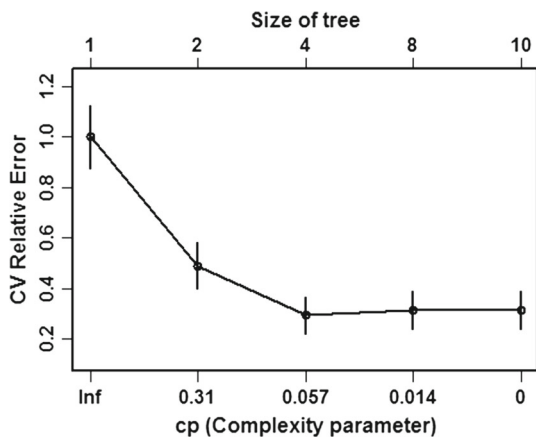




**Fig. 9** **a** Hierarchical Clustering of the data with class labels, **b** classical MDS projection of the data set in two dimensions



**Fig. 10** Self-organizing Map (SOM) projection



**Fig. 11** Error rate for different tree sizes for rpart method

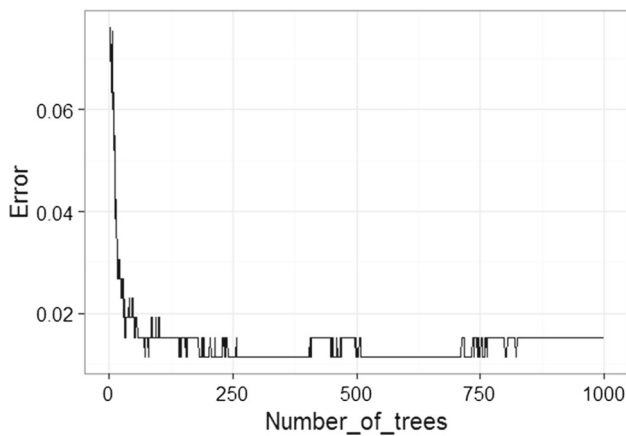
**(b) Random Forest**

Next, the Random Forest model is implemented to learn the data set. At first, the algorithm is trained with 1000 trees. Random forest algorithm uses bootstrap sampling (sampling

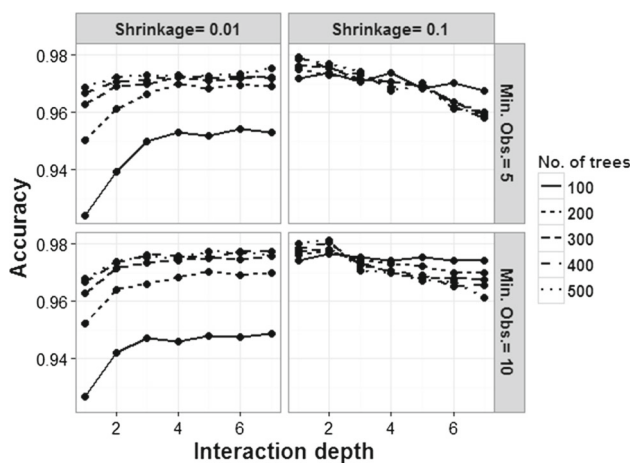
with replacement) of the data set and only a subset of the attributes are used for model building. For classification problems square root of the total number of predictor variables is the suggested value. So, the ‘mtry’ parameter was set at 6 (total number of predictors=40) in accordance with the author’s (Chen et al. 2004) suggestion. In the random forest algorithm 2/3<sup>rd</sup> of the data is used as training data for each tree and rest 1/3<sup>rd</sup> data is used to estimate the out-of-bag (OOB) error rate. In Fig. 12, OOB error is plotted against the number of trees. From the Fig. 12 we can see that the initially the OOB error rate is high and unstable but with increasing number of trees, this OOB error decreases and stabilizes closer to zero. So, for training of the final model 500 trees were used for the random forest learning.

**(c) Gradient boosting method**

Finally the gradient boosting using ‘gbm’ method was implemented to learn the dataset. Before going further different tuning parameters of the model needs to be discussed.



**Fig. 12** Error rate for different tree sizes for Random Forest



**Fig. 13** Accuracy for different set of tuning parameters for 'gbm'

Minimum number of observation denote the lower limit of the observations in the terminal node of any tree. Shrinkage is representative of the learning rate of the algorithm. Lower value signifies a model that learns slowly, while a higher shrinkage value may result in poor performance. Number of trees correspond to number of gradient boosting iterations. Higher number of trees fit the training data well, but too high value may lead to over-fitting. And, lastly interaction depth denotes the maximum number of splits on a tree. To tune these parameters a set of total 140 combinations were selected. Results of these runs are visualized in the Fig. 13. Based on these results optimal set of tuning parameters were chosen for final model building stage.

For the first case, all three types of MLF were not considered separately. No categorization were performed based on the type of MLF case. The first method to be applied is the decision tree method using 'rpart' package (Therneau et al. 2014) in R. After training the model, it was used to predict the results of testing data set to judge the predictive accuracy of the model. In Fig. 14, result of this model has been visualized using a decision tree. This decision tree suggests three

most important variables in classifying Non-MLF and MLF cases are minimum casting speed, minimum mould level and first ladle weight. This findings are in agreement with our previous findings. As no distinction was done between different MLF cases, and as most of MLF signals were due to slowdown, minimum casting speed of less than 1 results in MLF signal generation in 94 % of the cases. Other important variables are minimum mould level and first ladle weight. Decrease of metal level in the mould below a certain point also indicative of MLF signal. Lastly, the first ladle weight splits the remaining cases. This is actually in accordance with our earlier findings about true MLF cases, that happen during the ladle change operation. First ladle weight below a certain value indicative of the fact that at the start of casting for that slab ladle weight was already low, and a ladle change was impending.

Next two other learning algorithms—random forest (Liaw and Wiener 2002) and gradient boosting method (Ridgeway 2015) was used to train the model and after model building these models were validated on the unseen testing dataset. Comparison of different models is summarized in Table 1. Before delving further a brief description of the different performance metrics would be required. Accuracy is the proportion of the total cases that have been correctly classified, here both the positive and negative cases are considered simultaneously. While Sensitivity (also known as Recall) is the proportion of the positive cases (MLF cases in this case) that has been correctly identified by the model, Specificity denotes the proportion of the negative cases (no MLF cases) that has been correctly categorized by the classification algorithm. Positive predictive value (also known as Precision) is the ratio of the total number of true positive to the total number of predicted positive cases and similarly negative predictive value is the ratio of the true negative to the total number of predicted negative cases. Finally, for unbalanced class classification problem F1 score is one of the important evaluation metric. F1 score is representative of the improvement of the model compared to the baseline case. F1 score can be defined as,

$$F_1 = 2 \times \frac{(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})}$$

For highly unbalanced classes baseline case is considered where all the cases are predicted to belong to the majority class in the training data. So, for the baseline case F1 score will be zero as true positive cases are zero in this case. This suggests a higher F1 score signifies improvement over the baseline case.

As we can see for all the three methods accuracy achieved on the dataset is quite high, but this gives a false impression about the actual accuracy of the models, due to highly imbalanced nature of the dataset. Rather than the accuracy,

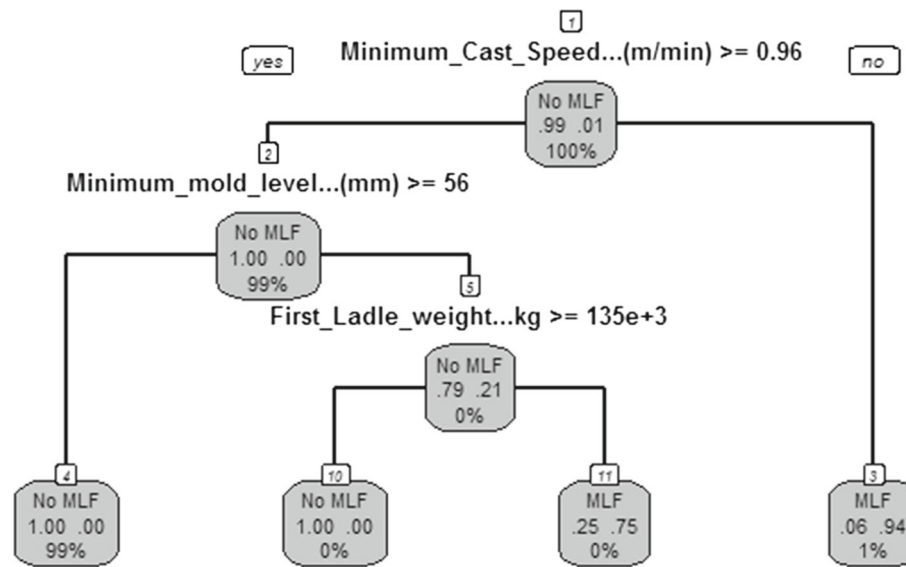


Fig. 14 Classification tree of non-MLF cases and MLF cases based on January, 2015 data

Table 1 Comparison of different models on the testing data set for Non-MLF and MLF case classification

Model	Sensitivity	Specificity	Pos Pred value	Neg Pred value	F1 Score	Accuracy
Rpart	0.80	0.998932	0.842105	0.998577	0.61	0.9975
Random Forest	0.80	0.999288	0.888889	0.998578	0.84	0.9979
Gradient boosting method	0.80	0.999288	0.888889	0.998578	0.84	0.9979
Combined model	0.80	0.999288	0.888889	0.998578	0.84	0.9979

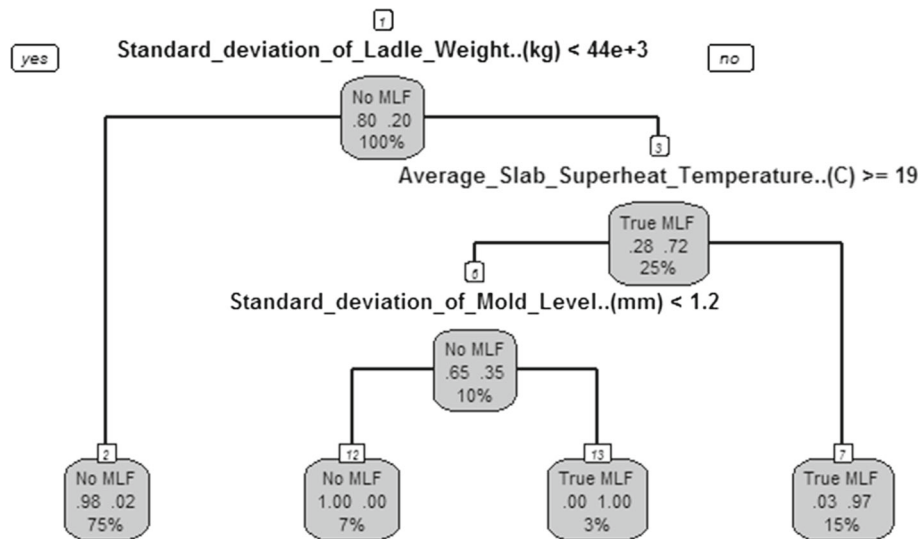
sensitivity and positive predictive values give us better picture about the accuracy of the models. Sensitivity is defined as the probability that the model will predict MLF given that it is an actual MLF case and positive predictive value is defined as the probability of actual MLF occurring when model has predicted MLF. According to the table, all three algorithms perform same in terms of sensitivity, while random forest and gradient boosting algorithms give us the best performance in terms of positive predictive value and accuracy.

Next, predictions from three of these models were used to build a combined predictive model. For the model stacking part results of the three different methods were used to predict MLF cases using generalized linear regression method. Results of the combined model improves on the results of rpart method, but it gives same accuracy as the random forest and gradient boosting algorithms. So, it suggests that for this particular problem random forest and gradient boosting algorithms performs the best, although there is scope for improvement.

For the next scenario, only true MLF cases are considered. Other two types of MLF cases were removed from the dataset. Next, this new dataset is divided in training and testing set. 70 % of the data was used as training set and rest was used as test data to validate the models.

The first method to be applied is the decision tree method using ‘rpart’ package. In Fig. 15, result of this model has been visualized using a decision tree. This decision tree suggests three most important variables in classifying Non-MLF and true MLF cases are standard deviation of the ladle weight, average slab superheat temperature and standard deviation of the mould level. Standard deviation of the ladle weight below a certain value separates 75 % of the data as Non-MLF cases with 98 % accuracy. Other important variables are average slab superheat temperature and standard deviation of the mould level. At the second level minimum superheat temperature below a certain limit separates 15 % of the total data set with 97 % accuracy, suggesting strong relation between superheat temperature and true MLF cases. Lastly, standard deviation of the mould level splits the remaining cases.

Next, two other learning algorithms – random forest and gradient boosting method (Ridgeway 2015) was used to train the model and after model building these models were validated using the testing dataset. Comparison of different models is summarized in Table 2. According to Table 2, gradient boosting algorithm gives best accuracy. Again, model stacking was done using these three methods and true MLF cases were predicted using generalized linear modelling. Results show that model blending improves the result com-



**Fig. 15** Classification tree of non-MLF cases and true MLF cases based on January, 2015 data

**Table 2** Comparison of different models on the testing data set for Non-MLF and true MLF case classification

Model	Sensitivity	Specificity	Pos Pred value	Neg Pred value	F1 Score	Accuracy
Rpart	0.7143	1.0000	1.0000	0.9375	0.83	0.9459
Random Forest	0.9048	1.0000	1.0000	0.9783	0.95	0.982
Gradient boosting method	0.9524	1.0000	1.0000	0.9890	0.98	0.991
Combined model	0.9524	1.0000	1.0000	0.9890	0.98	0.991

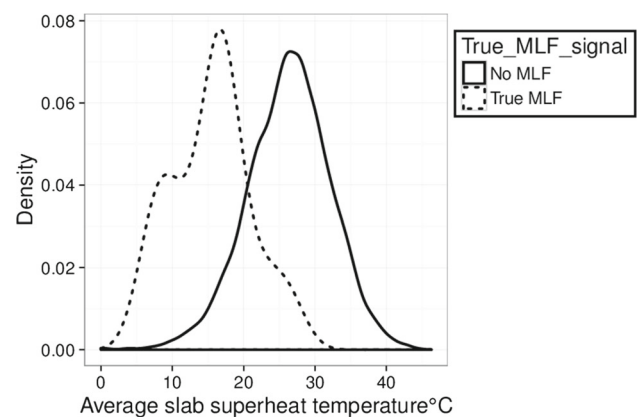
**Table 3** Results of hypothesis testing of superheat temperatures between MLF cases and true MLF cases

Variable	t-value	p-value	Lower limit	Upper limit
Average slab superheat temperature	8.289	3.789e−06	9.901647	17.024695
Minimum slab superheat temperature	6.5106	3.857e−05	6.786213	13.684915

pared to rpart method and the random forest algorithm, but fails to predict better than gradient boosting methods.

Based on the classification approach above, there seems to be strong linkage between superheat temperature and large scale mould level fluctuations during ladle change operation. To validate our intuition we conduct hypothesis test on average superheat temperature and minimum superheat temperature for two cases of non-MLF and true MLF. We assume a null hypothesis that for both the cases this two variables are not significantly different. Results from statistical test are summarized in Table 3. These results suggest there is significant difference in the values of average and minimum slab superheat temperature for non-MLF and true MLF cases.

In Fig. 16, difference in average superheat temperature has been visualized for both the cases of non MLF cases and true MLF cases. Shift of the mean to lower temperatures is quite evident for the true MLF cases from the Fig. 16. Also,



**Fig. 16** Comparison of average slab superheat temperature for no MLF and true MLF cases

the results of t-test substantiate the claim that there is significant difference in superheat temperature values for non MLF and true MLF cases. So, along with ladle change oper-



ation superheat temperature is one of the important process parameters influencing large scale mould level fluctuations. When Singh et al. (2012) numerically investigated the performance of different tundish furniture based on residence time parameters, they found that even small change in tundish temperature significantly affects the fluid flow in the tundish. In particular, non-isothermal consideration of the flow suggested buoyant effect can shift the centre of the circulatory flow in the tundish. This fluid flow feature in the tundish can significantly affect the flow of liquid metal in the mould as metal flow rate from tundish to mould is one of the important initial conditions to this phenomenon.

## Conclusion

In this study large scale mould level fluctuation in the thin slab caster at Tata Steel, Jamshedpur plant has been analysed using exploratory data analysis for knowledge discovery purpose about the phenomenon and predictive accuracy of the different classification algorithm in predicting mould level fluctuation phenomenon is also compared.

1. Large scale mould level fluctuations in thin slab caster can be classified in three distinct classes of—(a) mould level fluctuations due to slowdown, (b) lead heat mould level fluctuations and (c) mould level fluctuations during ladle change operations. MLF during ladle change operations are of prime importance for this study, as other two cases are not avoidable due to operational practices.
2. Exploratory data analysis suggests variation of casting speed, ladle weight (first and last) and variation of tundish weight as most influencing parameters in case of true MLF.
3. Classification of MLF (considering all three different cases) and non-MLF cases using different classification algorithms also substantiate our findings from the preliminary exploratory work. This part of the study suggests minimum casting speed, minimum mould level and first ladle weight as three most important variables in classifying Non-MLF and MLF cases. Comparison of different learning algorithms suggest random forest and gradient boosting as the ones with highest accuracy.
4. Next, classification of true MLF and non-MLF cases using different classification algorithms also confirms our claim from the preliminary exploratory study. In this case, ladle weight, average slab superheat temperature and standard deviation of the mould level are three most important variables in classifying Non-MLF and true MLF cases. Model comparison of different learning algorithms suggest gradient boosting as the one with highest accuracy.
5. Hypothesis test also suggest significant difference between average and minimum superheat temperature for non MLF cases and true MLF cases.
6. During ladle change operation tundish height decreases quickly as there is no incoming source of hot metal. At the same time low superheat results in decrease in fluidity of the molten metal in the mould. Both of which ultimately result in lower molten metal flow rate into the mould. That's why with higher casting speeds liquid level in the mould drops suddenly, before the mould level controller increases stopper rod opening to compensate for drop in liquid level in the mould. This phenomenon is characterised by typical signature of true MLF cases, where a sudden drop in metal level is noticed during ladle change operations.

**Acknowledgements** The authors are grateful to the TSCR Caster Automation, Electrical and Operations team, for their support and expert guidance on the subject matter which greatly assisted in this study.

## References

- Chen, C., Liaw, A., & Breiman, L. (2004). *Using Random Forest to learn imbalanced data*. Berkeley: Department of Statistics, University of California. 12.
- Colla, V., Matarese, N., & Nastasi, G. (2011). Prediction of under pickling defects on steel strip surface. *International Journal of Soft Computing and Software Engineering*, 1(1), 9–17. doi:10.7321/jscse.v1.n1.2.
- de Beer, P. G., & Craig, K. J. (2008). Continuous cast width control using a data mining approach. *Ironmaking & Steelmaking*, 35(3), 213–220. doi:10.1179/030192307X233052.
- Furtmueller, C., del Re, L., Bramerdorfer, H., & Moerwald, K. (2005). Periodic disturbance suppression in a steel plant with unstable internal feedback and delay. In *Proceedings of 5th international conference on technology and automation, ICTA* (vol. 5, pp. 1–6).
- Galili, T. (2015). dendextend: an R package for visualizing, adjusting, and comparing trees of hierarchical clustering. *Bioinformatics (Oxford, England)*, 31(22), 3718–3720. doi:10.1093/bioinformatics/btv428.
- Girase, N. U., Basu, S., & Choudhary, S. K. (2007). Development of indices for quantification of nozzle clogging during continuous slab casting. *Ironmaking & Steelmaking*, 34(6), 506–512. doi:10.1179/174328107X168075.
- Gursoy, K. A. (2014). *Quantifying the effect of flow rate controllers on liquid steel flow in continuous casting mold using CFD modeling*. Middle East Technical University.
- Hajari, A., & Meratian, M. (2010). Surface turbulence in a physical model of a steel thin slab continuous caster. *International Journal of Minerals, Metallurgy, and Materials*, 17(6), 697–703. doi:10.1007/s12613-010-0376-7.
- Javurek, M. (2008). Lingotamento Continuo-Continuous Casting.png. [https://commons.wikimedia.org/wiki/File:Lingotamento\\_Continuo-Continuous\\_Casting.png](https://commons.wikimedia.org/wiki/File:Lingotamento_Continuo-Continuous_Casting.png).
- Kant, S., Jha, P. K., & Kumar, P. (2013). Investigation of effect of dam on intermixing during ladle changeover in six strand billet caster tundish. *Ironmaking & Steelmaking*. <http://www.tandfonline.com/doi/abs/10.1179/1743281211Y.000000007#.VtQQ6PI97cc>. Accessed 29 February 2016.

- Kumar, D. S., Rajendra, T., Sarkar, A., Karande, A. K., & Yadav, U. S. (2007). Slab quality improvement by controlling mould fluid flow. *Ironmaking & Steelmaking*, 34(2), 185–191. doi:10.1179/174328107X155330.
- Lee, P. D., Ramirez-Lopez, P. E., Mills, K. C., & Santillana, B. (2012). Review: The “butterfly effect” in continuous casting. *Ironmaking & Steelmaking*, 39(4), 244–253. doi:10.1179/0301923312Z.00000000062.
- Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, 2(3), 18–22. <http://cran.r-project.org/doc/Rnews/>
- Liu, R., Sengupta, J., Crosbie, D., Yavuz, M. M., & Thomas, B. G. (2011). Effects of stopper rod movement on mold fluid flow at ArcelorMittal Dofasco’s No. 1 continuous Caster AISTech 2011 proceedings (vol I(1), pp. 1619–1631).
- Liu, R., Thomas, B. G., Kalra, L., Bhattacharya, T., & Dasgupta, A. (2013). Slidegate dithering effects on transient flow and mold level fluctuations. In *AISTech 2013 proceedings* (pp. 1351–1364).
- Liu, R., Thomas, B. G., & Sengupta, J. (2012). Simulation of transient fluid flow in mold region during steel continuous casting. *IOP Conference Series: Materials Science and Engineering*, 33, 012015. doi:10.1088/1757-899X/33/1/012015.
- Matsumiya, T. (2006). Recent topics of research and development in continuous casting. *ISIJ International*, 46(12), 1800–1804. doi:10.2355/isijinternational.46.1800.
- Nastasi, G., Colla, V., Cateni, S., & Campigli, S. (2016). Implementation and comparison of algorithms for multi-objective optimization based on genetic algorithms applied to the management of an automated warehouse. *Journal of Intelligent Manufacturing*. doi:10.1007/s10845-016-1198-x.
- Ordieres-Meré, J., Martínez-de-Pisón-Ascacibar, F. J., González-Marcos, A., & Ortiz-Marcos, I. (2008). Comparison of models created for the prediction of the mechanical properties of galvanized steel coils. *Journal of Intelligent Manufacturing*, 21(4), 403–421. doi:10.1007/s10845-008-0189-y.
- R Core Team. (2014). R: A language and environment for statistical computing. Vienna, Austria. <http://www.r-project.org/>
- Rackers, K. G. & Thomas, B. G. (1995). Clogging in continuous casting nozzles. In: *78th steelmaking conference proceedings* (vol. 78, pp. 723–734).
- Ridgeway, G. (2015). gbm: Generalized boosted regression models. <https://cran.r-project.org/package=gbm>
- Sharafi, S. M., & Esmaeily, H. R. (2010). Applying data mining methods to predict defects on steel surface. *Journal of Theoretical & Applied Information Technology*, 20, 87–92.
- Shaver, J. W. (2002). *Measurement of metal/slag interfacial phenomena in thin slab caster*. Retrieved from [http://ccc.illinois.edu/PDFFiles/Theses/2002\\_SHAVER\\_Joseph\\_MSThesis.pdf](http://ccc.illinois.edu/PDFFiles/Theses/2002_SHAVER_Joseph_MSThesis.pdf).
- Singh, A. K., Pardeshi, R., & Goyal, S. (2011). Integrated modeling of tundish and continuous caster to meet quality requirements of cast steels. In *Proceedings of the first world congress on integrated computational materials engineering* (vol. 1, pp. 81–85).
- Singh, V., Ajmani, S. K., Pal, A. R., Singh, S. K., & Denys, M. B. (2012). Single strand continuous caster tundish furniture comparison for optimal performance. *Ironmaking & Steelmaking*, 39(3), 171–179. doi:10.1179/1743281211Y.0000000065.
- Suzuki, D. (2004). Formulation of mold level control model by molten steel flow analysis method. *Nippon Steel Technical Report*, 89, 46–49.
- Tang, L., Liu, W., & Liu, J. (2005). A neural network model and algorithm for the hybrid flow shop scheduling problem in a dynamic environment. *Journal of Intelligent Manufacturing*, 16(3), 361–370. doi:10.1007/s10845-005-7029-0.
- Therneau, T., Atkinson, B., & Ripley, B. (2014). rpart: Recursive partitioning and regression trees. <http://cran.r-project.org/package=rpart>
- Thomas, B. G. (2001). Continuous casting of steel. *Modelling for Casting and Solidification Processing*. doi:10.1049/sqj.1963.0042.
- Thomas, B. G. (2013). *Fluid flow in the mold* (pp. 1–41). <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.503.5487&rep=rep1&type=pdf>.
- Torgo, L. (2010). *Data mining with R, learning with case studies*. London: Chapman and Hall/CRC. <http://www.dcc.fc.up.pt/~ltorgo/DataMiningWithR>
- Wang, H., & Wu, K. (2003). Modeling and analysis for multi-period, multi-product and multi-resource production scheduling. *Journal of Intelligent Manufacturing*, 14(3), 297–309. doi:10.1023/A:1024645608673.
- Wehrens, R., & Buydens, L. M. C. (2007). Self- and super-organizing maps in R?: The kohonen package. *Journal of Statistical Software*, 21(5), 1–19. doi:10.18637/jss.v021.i05.
- Zhang, L., Yang, S., Cai, K., Li, J., Wan, X., & Thomas, B. G. (2007). Investigation of fluid flow and steel cleanliness in the continuous casting strand. *Metallurgical and Materials Transactions B: Process Metallurgy and Materials Processing Science*, 38(February), 63–83. doi:10.1007/s11663-006-9007-0.