**Error analysis**:

1. Operand count not sufficient :

Analysis of results show that the false identification of displayed equations as embedded ones are sometimes due to the presence of more than one word blob as operands to one operator.

Consider fig. cbook16.png, the first 3 chemical equations are classified correctly. But the last equation after removing the equation number has 8 operands but 3 operators. So, # of operands >= 2\*# of operators (see fig: wordBlobCbook.png).

1. Failure to understand the context:

In some cases if the operands of a chemical equation have alkyl or halide group, they denote them as R and X respectively. But these symbols are not in the periodic table. Hence when the substrings from the OCR output are searched in the dictionary, it comes back negative. And the equation is detected as a non-chemical one. See 4th equation bounded with a green rectangle of fig: cbook16.png.

1. Single character detection:

Some chemical equations have some reactants just on top of the arrow and they enter the bounding box of the arrow. In those cases when we extract only single characters from the word blob information, we don’t get the arrow as it is no longer treated as a single character within its bounding box. Here the operator, arrow doesn’t get detected and # of operands: # of operands operators exceeds 2. So, this is segmented as a embedded equation. See. Fig. escan3.png

1. False detection of embedded equations:

In a few cases, text lines consisting of embedded expressions are identified as displayed equations (see, fig. chem31.png), but in those cases the text lines contain more mathematics than normal text. However, identification of mathematics intensive text lines as displayed equation is not a sever error.

**Conclusion:**

We have presented an automated chemical and non-chemical equation segmentation system able to detect displayed equations of both categories from heterogeneous document images. The method is based on detecting operators as +,-, -> which are common in both chemical and non-chemical equations, segmenting the displayed equations and then running them through an OCR to classify. A publicly available database for mathematical documents and scanned images of various chemistry books of both English and Bengali language are used in this study. The experimental results demonstrate the efficiency of our proposed method.

This present study points to several new research avenues to be explored further. The over-all performance itself demands further research in this area. Excessive degradation due to aging and improper digitization of the documents images give rise to broken and merged characters which give conflicted output from the OCR. In future, design of a better integrated OCR specifically for chemical equations would be ventured in. In addition, formation of electron bond matrix of a chemical compound in a reaction mentioned in the introduction section would be a high-level goal in the future.