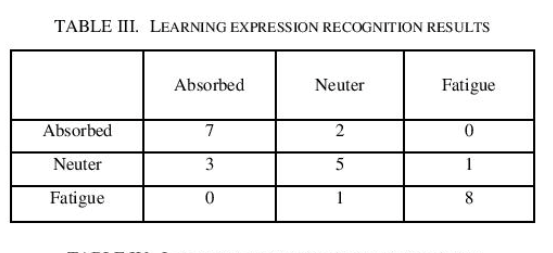
CẢI TIẾN OCR

Những thứ chưa làm được hoặc làm được nhưng kết quả chưa tốt cần cải tiến:

**1, OCR ảnh nhỏ cắt ra có độ phân giải thấp chưa đọc được:**

- cần thực hiện xử lí ảnh nhỏ sau khi cắt ra như tăng sáng, khử nhiễu, tăng kích thước ảnh,…

**2, OCR bảng:**

- Nguyên lý của OCR table mà nhóm đang thực hiện đó là dò tìm các đường của bảng sau đó cắt ra thành các ô nhỏ trong bảng

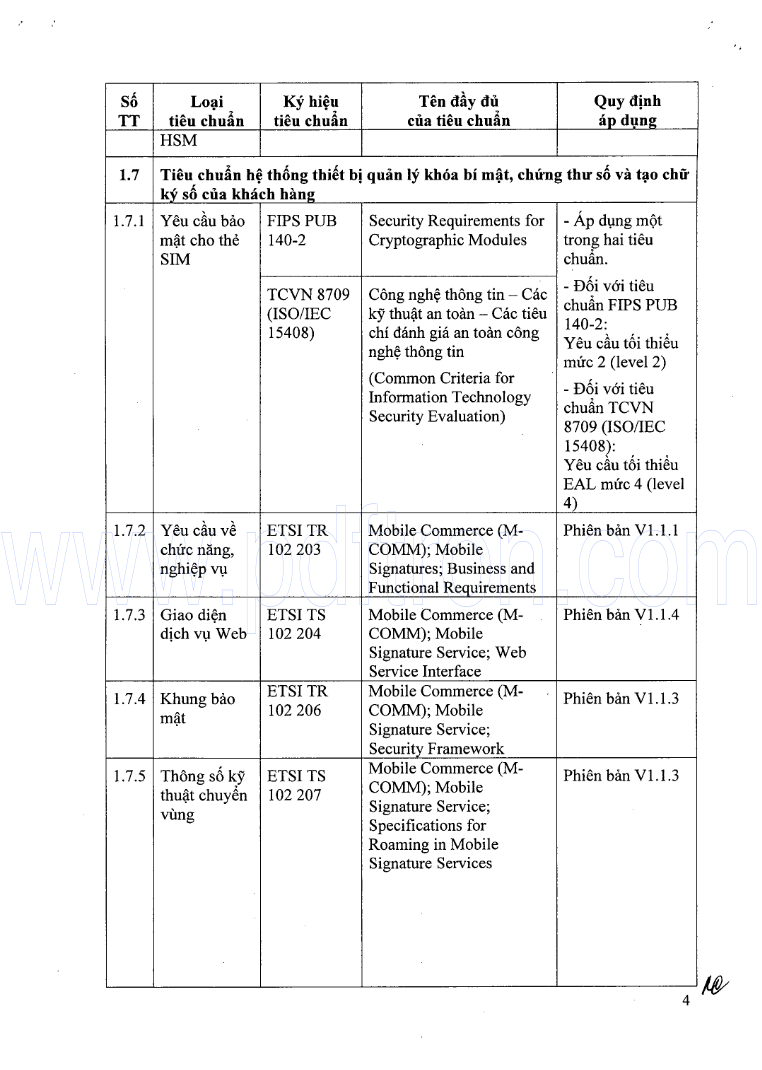
*Hình 1*

+ Như hình trên sau khi thực hiện dò tìm các đường viền, cạnh của bảng sẽ thu được 16 ô nhỏ

+ Sau đó, tiến hành đọc OCR từng ô nhỏ rồi xuất ra thông tin qua file excel theo thứ tự tương ứng.

=> Vấn đề đặt ra mà nhóm chưa giải quyết được:

+ Nếu như bảng (table) có các ô đều như hình trên (*hình 1)* thì khi xuất ra file excel sẽ cho kết quả tốt, độ chính xác cao.

 + Nhưng với trường hợp trong bảng mà các ô không đều (một số ô bị gộp) như hình bên dưới (*Hình 2)* thì lại cho kết quả chưa chính xác vị trí các ô tương ứng khi xuất ra excel.

*Hình 2*

=> Cần cải tiến cho trưởng hợp này.

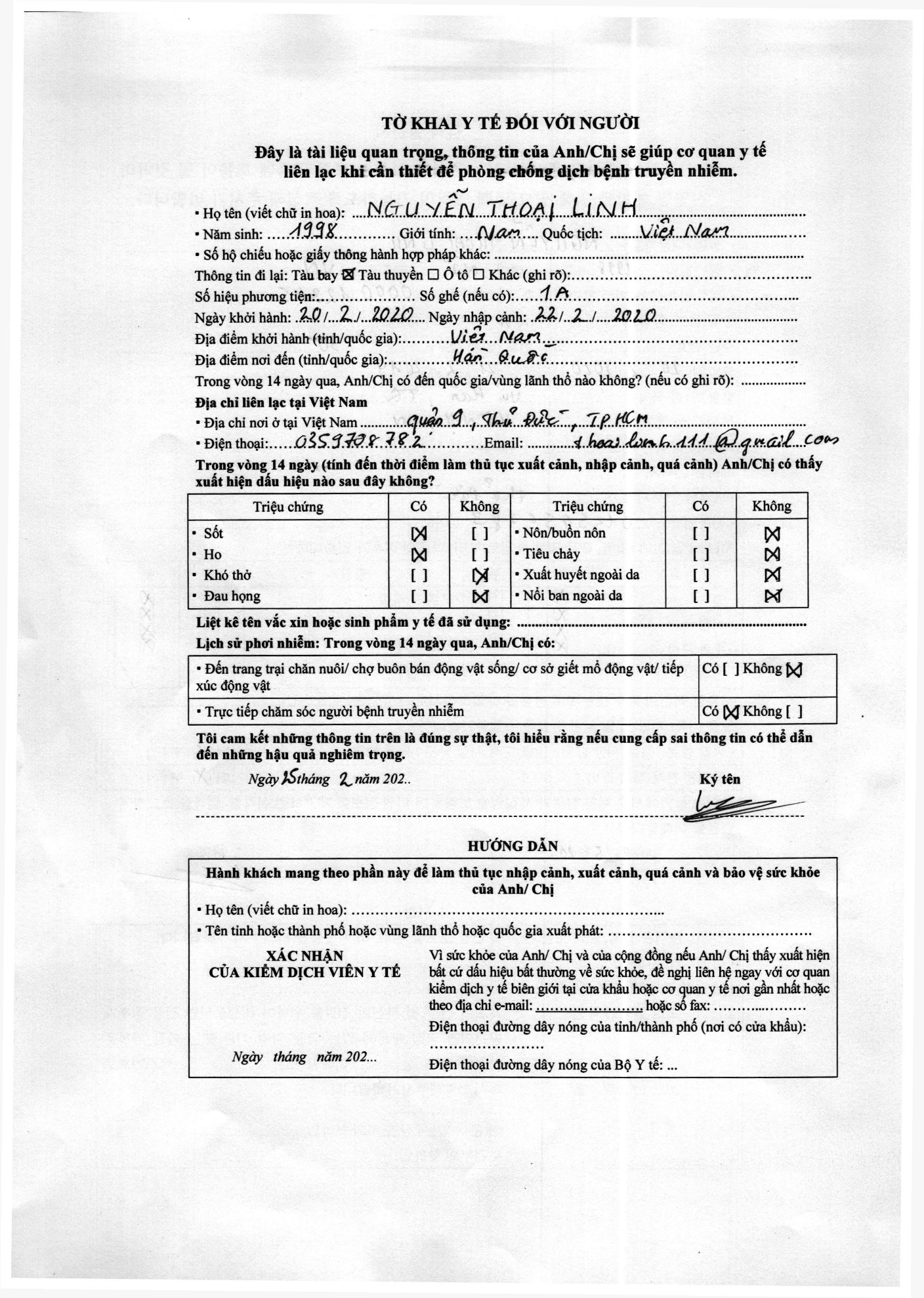
**3, OCR biểu đồ**

Biểu đồ có rất nhiều loại khác nhau như biểu đồ cột, biểu đồ hình tròn, biều đồ đường,… với nhiều thông số khác nhau xuất hiện ở nhiều vị trị khác nhau.

Vấn đề này nhóm chưa giải quyết được.

**4, Nhận dạng chữ viết tay.**

- Trong việc nhận diện văn bản, trong thực tế thì có cả loại văn bản có form sẵn rồi cho người sử dụng viết tay điền vào. Việc OCR (như đã nêu ở trên) thì chỉ đọc được các ký tự quang học, vậy nên để đọc được các chữ viết tay, cần phải xây dựng thêm tính năng nhận diện chữ viết tay.



*Hình 3*

Vấn đề đặt ra là các dạng form để điền vào có nhiều dạng khác nhau -> chữ viết tay điền vào nằm ở nhiều vị trí khác nhau trong văn bản

=> cần phải train để chương trình nhận diện được các loại form mẫu để việc nhận diện chữ viết tay được dễ dàng hơn

**5, Áp dụng xử lý ngôn ngữ tự nhiên(NLP) :**

- Sau khi đọc văn bản, tất nhiên sẽ không thể chính xác 100% được. Sẽ có các trường hợp đọc sai về chính tả lẫn nghĩa. Ví dụ: từ “Cộng hòa” được nhận diện thành “Cộng hỏa” => cần áp dụng NLP để tự động sử thành “Cộng hòa”

- Mẫu điền form viết tay có cả trường hợp input đầu vào là chữ nước ngoài, nhiều ngôn ngữ khác nhau => cần cải tiến, bổ sung thêm vấn đề này.